UNIVERSITAT DE BARCELONA

Machine-Learning Applied Methods

Sebastián Mauricio Palacio

ERSITAT DE
CELONA

2020

PhD in Economics | Sebastián Mauricio Palacio

UNIVERSITAT DE
BARCELONA

PhD in Economics

Machine-Learning Applied Methods

Sebastián Mauricio Palacio

LIBERTAS PERFVNDET
OMNIA LVCE

UNIVE
BARC

PhD in Economics

Thesis title:

Machine-Learning Applied Methods

PhD student:
Sebastián Mauricio Palacio

Advisor:
Joan-Ramon Borrell

Date:
February 2020

UNIVERSITAT DE BARCELONA

To Kine, Marcela and Antonia

Acknowledgments

I was accepted to the PhD program at the Department of Econometrics, Statistics and Applied Economics in May 2016. A lot of things have happened since that day and many people have played a part in finally presenting my thesis.

I would first like to thank to my supervisor Joan-Ramon Borrell and my tutor Montserrat Guillen, whose invaluable contributions and guidance encouraged me never to give up. Ever since I did my Master's degree, Joan-Ramon has had an open mind and given me the freedom and trust to follow my own path. For this I am grateful. Also for all the encouragement and the precise comments on my work, and for always taking time to answer my numerous questions whether related to the thesis or not. Thank you both!

I would like to acknowledge Zurich Insurance Ltd for their financial support and for letting me to conduct my research in their installations and providing me not only with technological and data resources, but also giving me total freedom and flexibility. In particular, I owe gratitude to my colleagues in the Advanced Analytics Department. Especially, I would like to thank to Cristina Rata for her trust and support, and Gledys Alexandra Sulbarán Goyo for her companionship in the workplace since my first day at Zurich.

As well, I would like to say thank you to all the members of the Public Policy Section of the Universitat de Barcelona. From the start, I felt the friendly and welcoming atmosphere in the department. The doors were always open and everybody was always eager to help or give advice in time of need.

From UB I would like to thank Stephan Joseph and Francisco Robles both for interesting academic discussions but also for all the good times in Barcelona. I would also like to thank my department colleagues: Carlitos, Max, Paula and Tania for being amazing persons and for always being there with a smile.

I would also like to thank to UBEconomics, especially Jordi Roca who has helped me with administrative issues from my first day as a PhD student.

Although the people mentioned above are very important in the process of finishing this thesis, I would never have been able to finish it without the support from family and friends. I wish to thank my family for always supporting and believing in me, who made all of this possible, especially mum and my grandma. To my dear friends: Diego, Germán, Emanuel and Nacho. I really appreciate our friendship. I would never forget the day you traveled thousands of kilometers just to be with me.

Last but not least, to Kine Bjerke, my wife and the love of my life. You not only moved far away from your family because of me but as well you have been by my side throughout this entire process. I really do not know how all of this would have been possible without you. There are not word that can express how much I

appreciate that you chose to spend your life with me.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Prediction versus Causal Inference

Causality and impact policy evaluation researchers have a long tradition in economics. However, as pointed out by Kleinberg et al. (2015), there are many economic applications where causal inference is not central, but instead where prediction may be more suitable. Recently, several economist authors started paying attention to prediction economy problems. Subfields of economics such as crime policy (Chandler et al., 2011; Berk, 2012; Goel et al., 2016), political economy (Grimmer and Stewart, 2013; Kang et al., 2013), insurance and risk economics (Bjorkegren and Grissen, 2018; Ascarza, 2018), public-sector resource allocations (Naik et al.; 2016; Engstrom et al., 2016), wealth economics (Blumenstock et al., 2015; Jean et al., 2016; Glaeser et al., 2016), energy economics (Yu et al., 2008; Yu et al., 2014; Afkhami et al., 2017; Chen et al., 2018), etc., are nowadays common examples.

Several factors have contributed to rethinking the way in which empiricists evaluate economic problems: First, data is notably expanding with new technologies (90% of the data today was created in the last two years -see Figure 1.1-); second, private and public sector are continuously increasing the amount and quality of collected data (structured and unstructured data-see Table 1.1); third, with the increasing use of machine learning and deep learning techniques, we can now exploit large data-sets and find much more complex patterns.

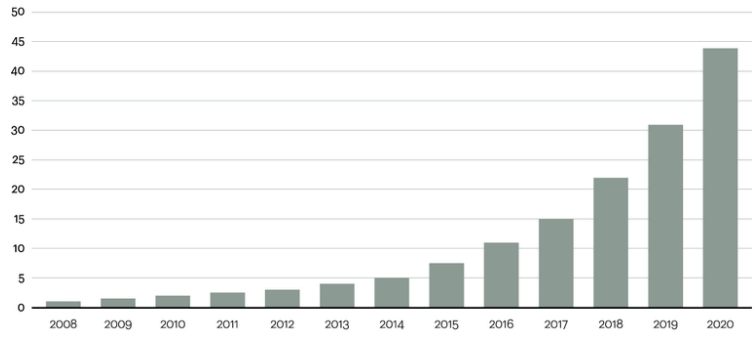


Figure 1.1. Data Volume Growth by Year in Zettabytes.
source: Hammad et al. (2015)

Platform	**Measure**	**2018**
Netflix	Users stream per Hrs	97,222
Youtube	Videos watched	4,333,560
Twitter	Tweets	473,400
Skype	Calls	176,220
Instagram	Photo posts	49,380
Spotify	Songs streamed	750,000
Google	Searches	3,877,140
Internet	Use of data (GB)	3,138,420

Table 1.1. Data Generated per Minute.
Note: Table Based on Data Never Sleep 6.0 by Domo Platform

Traditionally, statistic models in economics are almost used exclusively for causal inference where a common assumption is that models that posses high explanatory power naturally posses high predictive power. However, not distinguishing between prediction and causal inference, has a large impact on the statistical assumptions and on its implications. In predicting problems, the focus is on methods that enhance prediction capabilities as opposed to the assessment of marginal effects on target variables. Moreover, they take into account the importance of performance in terms of out-of-sample errors.

But, why causal inference and predicting are conceptually different? Causal inference tries to answer the question of how much underlying factors affect a dependent variable, that is, they try to test causal hypotheses that emerge from theoretical models. Contrarily, prediction models are solving fundamentally different problems compared with much of the empirical work in economics. Predictive models seek to predict new observations, that is, predict new values of the dependent variable

given their new values of input variables. In short, the difference arises because the goal in causal inference is to match the statistical model and the theoretical model as closely as possible. In contrast, in predictive modeling the statistical model is used as a tool for generating good predictions of a target variable.

This disparity can also be expressed mathematically. For instance, Mean Squared Error (MSE) of a new point x_0 can be decomposed in the following way (Hastie et al., 2009):

$$MSE(x_0) = Bias^2 + Var(\hat{f}(x_0)) + \sigma^2$$

Bias is the result of misspecifying the statistical model f. $Var(\hat{f}(x_0))$ is the variance which is the result of using samples to estimate the statistical model. σ^2 is the error term that exists even if the model is correctly specified. This formulation reveals the disparity between causal inference and prediction. The former focuses on minimizing the bias (Ordinary least square, for instance, is the best linear unbiased estimator), i.e to get the most accurate representation of the underlying theory. However, if we ensure zero bias, we cannot trade bias for variance reduction. In spite of causal inference, predictive modeling exploits empirically this trade-off to get best performance in out-sample data. It seeks to minimize the combination of bias and variance, occasionally sacrificing interpretability.

These several differences are summarized in Table 1.2.

	Causal Inference	**Prediction**
Definition	Causation: The statistic model represents a causal function.	Association: The statistic model seek the association between target variables and predictors.
Foundation	The statistic model is constructed based on a theoretical model that tries to assess the relation between a dependent variable and explicative variables and to test the causal hypothesis.	The statistic model is a representation from the data. Interpretability is usually not required.
Focus	To test already existing hypotheses.	To predict new observations.
Target	Bias	Variance

Table 1.2. Difference Between Causal Inference and Predictive Models

In the end, ML and traditional statistical models have different goals. Traditional econometric models are based on theoretical foundations which are mathematically proven. However, they require that the input data satisfies strong assumptions (like random sampling observations, perfect collinearity, etc.) and a particular distribution for the error term. On the other hand, ML and DL are data driven models, and the only assumptions that we make are that observations are independent (which

sometimes is not even necessary) and that the training and test set follows the same joint distribution. Additionally, as ML and DL are focused on out-sample performance, it is expected that they get better results in terms of prediction. On the other hand, statistical models focus on goodness-of-fit (discrepancy between observed values and the values expected under the model), i.e. bias, which increase complexity and may decrease predictive power. Therefore, we expect that by combining those techniques, we can compensate the disadvantages of both and we can enhance and robust our results.

1.2 Why Machine Learning and Deep Learning?

Machine Learning (ML) and Deep Learning (DL) techniques are particularly effective in predicting. This branch of computer science, which gained popularity in the 1980s, has recently been successfully employed in several fields thanks to a number of technological advances. Basically they are algorithms that are used for prediction, classification and clustering with structured and unstructured data (see Varian, 2014 and Athey, 2017 for an overview of some of the most popular methods). These, combined with the development of new and more efficient programming languages, have drastically reduced the computational time. We favor the application of ML and DL algorithms over more traditional techniques simply because most empirical approaches are not accurate enough in their forecasts (Zhang and Xie, 2008; Kleinberg et al., 2015; Zhao et al., 2018). Prediction problems should be more exclusively focused on the target variable and on the accuracy of the prediction, and not on dependent variables and their causal effect. They also strive to obtain better performance in the measurement error. In this sense, traditional econometric methods are not optimal, given that they focus mainly on unbiasedness. When it comes to prediction, therefore, they tend to be "overfitted" (i.e., fitted too closely to a particular data-set), and, in consequence, to generalize poorly to new, unseen data.

As we explained, ensuring unbiasedness in in-sample error allows no trade-off with variance reduction. In spite of traditional econometric techniques, ML and DL fully exploit the possibility of this trade-off to get the best performance in out-sample data. By focusing on prediction problems, ML and DL models can minimize forecasting error by trading off bias and variance.

It follows, therefore, that ML and DL algorithms are specifically designed for making predictions. Moreover, ML and DL are able to exploit several data types and complexities. But perhaps their main advantage is the fact that computers can be programmed to learn from data, revealing previously hidden findings as they

discover historical relationships and trends. ML and DL techniques can improve the accuracy of predictions by removing noise and by taking into account many types of estimations, although not necessarily without bias. Moreover, ML and DL allows for a wide range of data, even when we have more predictors than observations, and it admits almost every type of functional form when using decision trees, ensuring a large interaction depth between variables. Of course, the downside of ML and DL techniques is biased coefficients; however, if our main concern is the accuracy of the prediction, then any concern regarding biased estimators becomes almost irrelevant.

1.3 Structure and Objectives of this Thesis

Beyond the advantages of ML and DL techniques, the main relevant objective here is to aboard empirically several economic prediction problems. Our empirical application in the next chapters highlights how improved prediction using machine learning and deep learning techniques can have large impacts on economics compared to traditional econometric techniques.

Although we center on three main economic fields (transport, energy and insurance), the cases we introduce are real world cases regarding traditional and well-known economic problems, and they aboard a large variety of sub-fields (from supervised, semi-supervised, unsupervised learning, to complex deep learning and time series models). This lets us obtain a better knowledge on how traditional economic problems can be enriched by those techniques.

The thesis is structured in six chapters of which introduction is the first. Chapter 2 **"Predicting Collusive Patterns in Electricity Markets: Case of Liberalized Markets with Regulated Long-Term Tariffs"** tries to shed light on the question of how mandated auctions affect liberalized electricity markets and to what extent collusion and price volatility can be accountable for price increments. Thereby, we examine the Spanish electricity market and the introduction of fixed-price forward contract obligations implemented between 2007 and 2013. The last auction was held on December 2013, however, the next day, the energy regulator declared it invalid. Although the final auction price was 7 percent higher than the daily price, the regulator explained that the causes were essentially exogenous to the firms and no penalties were imposed. An arising question is however, whether this result was extraordinary at all but rather a repeated hidden action. Thereby, we seek to validate the hypothesis of the existence of strong incentives to increase prices in daily electricity markets when fixed long-term tariffs are applied. To do so, we seek answers to the following questions: Do regulated long-term tariff auctions trigger collusion in daily markets, i.e. will companies try to influence prices expectations

in daily markets to get better deals in the auction market? Furthermore, to what extent could they be affecting daily prices?

Respectively, predicting the collusive phases in price is the first contribution of this dissertation. One of our major results is that prices increased by 15 percent 70 days before the mandated auctions. This result is contrary to what literature to date has been claimed, that the introduction of such tariffs increases competition and leads to supply prices that are closer to the marginal costs.

Strong insight of the collusive behavior are derived from a theoretical model: First, the inherent characteristics of markets of this type serve as an incentive to collusion. This is supported by the Spanish electricity market characteristics: a high concentration of generation capacity and a low level of interconnectivity. Second, possible exogenous price shocks generate a perfect subgame equilibrium where prices are higher than without them, which is very related to the liberalized process in the Spanish market. A deficient design (repeated auctions and fixed prices) in an environment of natural concentration and high price volatility seems to be the main reasons why firms colluded.

The third chapter **"Machine Learning Forecasts of Public Transport Demand: A comparative analysis of supervised algorithms using smart card data"** contributes towards predicting public transport demand using smart card data and understating how it is affected by nominal increases in fares.

Public transport in the Autonomous City of Buenos Aires (the capital city of Argentina) is provided in an integrated system that combines urban buses with suburban buses, an incipient underground metro network and inter-city trains. Passengers use a smart card (SUBE card) which provides extremely rich and reliable source of data. In the analysis period, bus fares suffered three different nominal increases, which gives us a unique opportunity to evaluate not only interpretability and predictive power but also demand elasticity. Thus, chapter three presents various supervised machine learning and linear model estimations which use smart card data in order to compare predictive power, interpretability and demand elasticities. Given the obtained results from the empirical exercise, it seems that supervised machine learning algorithms are much more accurate than linear models for predicting demand. Second, both type of models show very similar outcomes: Time variables, cross elasticities and weather precipitations are the most influential variables in predicting public transport demand. One particularly notable outcome is that none of the supervised algorithms showed responsiveness to nominal fare increases (We have evaluated this formulation during a period where nominal fares increased around 80 percent and real fares did not change at all). Contrarily, our lineal model specification showed a demand elasticity of -0.31, with an initial shock of -0.47, supporting the hypothesis of a money illusion effect.

Chapter 4 **"Abnormal Pattern Prediction in the Insurance Market: Fraudulent Property Claims"** addresses a well-known predictive problem in insurance markets but which is also very difficult to aboard because of its nature: Fraud. It has been estimated that fraud cases represent up to 10 percent of all claims in Europe (€204 billion claimed cost) and account for around 10-19 percent of the payout bill (The Impact of Insurance Fraud, 2013).

In practice, fraud detection prediction problems are characterized by the simultaneous presence of skewed data (Phua et al. -2010- find that more than 80 percent of a review of 10 years of fraud detection studies have a percentage of fraud cases below 30 percent), a large number of unlabeled data (information available is usually only related to investigated cases) and a dynamic and changing pattern.

In this chapter, we propose a methodology based on semi-supervised techniques and we introduce a new metric – the Cluster Score- for fraud detection which can deal with these practical challenges. To represent this case, we draw on information provided by a leading insurance company. Particularly, we seek to predict fraudulent property claims which has been largely neglected by the fraud insurance literature.

Out of a total of 303,166 property claims submitted between January 2015 and January 2017, only 7,000 cases were investigated by the Investigation Office (IO). Of these, only 2,641 were actually true positives (0.8 percent from the total). This means, we do not know which class the remaining cases belong to. Our main results of the proposed methodology reveals that we are able to predict 97 percent of the total cases. However, the added value depends on the fraud cases that were never investigated (because they were not considered as suspicious cases) and are predicted as fraudulent. We, therefore, randomly set aside 10 percent of the data (30,317 claims). Of these, we were able not only to predict the total fraud cases (271 claims) but, additionally, 367 non-investigated cases were predicted as fraudulent. Those cases were sent to the IO for analysis, which 333 were found to present a very high probability of being fraudulent. In short, we managed to increase the efficiency of fraud detection by 122.8%.

Lastly, the fifth chapter **"Risk Categorization and Self-Reported Mechanisms in Automobile Insurance Markets"** also presents a well known asymmetry information problem: Before a contract can be signed, insurance companies know next to nothing about their potential new customers, while the latter tend to underreport prior claims when switching to a new company. Basic insurance theory suggests that risky customers will not reveal their true nature, and therefore, a suboptimal Pareto equilibrium with an average premium will be reached (Arrow, 1963; Akerloff, 1970). However, the first questions we seek to address is: Are always all "bad risks" pretending to be "good risks" as theory suggests? Or is it more nuanced,

in that only a subset misreport their history?

In this chapter, using past performance shared data from representative insurers in Spain, we test the hypothesis that not all high risk individuals pretend to be low risk. Based on this hypothesis, we combine self-reported data and observable characteristics from potential customers to enhance predictive power of risk classification, to identify the main features that drive misreporting and to find the most important variables for predicting risk.

Especially, we propose using a deep variational autoencoder (VAE) model and the match between internal customer data and potential customers data. We then approximate the risk by employing clusters as input variables. The VAE model not only allows us to reduce the large number of variables to their true nature but it can also be transformed into a powerful outlier model. With this methodology, we are able to predict (ex-ante) 80-87% of the risky customers. However, we fail to predict risky customers if we do not combine self-reported data and the observable characteristics. This result is supported by a theoretical model which states that, combining self-reported and risk categorization mechanisms, a private monopoly can get higher profits than the classic private information equilibrium.

Additionally, we find that the most important variables for measuring risk are not related to self-reported prior claims but rather to self-reported years as insured. In addition, cluster constructed variables related to the customers' zip code and customer characteristics were very significant. Similarly, the following were also found to be systematically important variables: if the insured was the owner and first driver in the policy, if the customer's age was higher than 65, if the insured was male or female and the number of license years.

Lastly, Chapter 6 concludes the thesis with a presentation of the main results from the previous chapters and provides some insights of the predictive models in the economic field.

2 Predicting Collusive Patterns in Electricity Markets: Case of Liberalized Markets with Regulated Long-Term Tariffs

2.1 Introduction[1]

When considering liberalized electricity markets, we assume the existence of strong competition between firms that leads to lower consumer prices. However, it is well known that electricity generation markets tend towards natural concentration due to their structural characteristics (few participants, transparent information, frequent interaction, high sunk costs and high market-shares). Moreover, liberalized markets often face high price volatility, derived from their instantaneous nature (dependency on international prices, real time demand, wind, non-availability, and interconnectivity restrictions). This tends to have a negative effect on the price consumers pay resulting in unpredictable and higher tariffs. Regulation may, therefore, represent a tool that can help provide an essential and, what is more, an affordable service. Long-term contract auctions are a widely used mechanism in many deregulated markets. In these, regulators mandate auctions where suppliers can make offers to provide an amount of energy at a fixed price during a set period of time. The main objective of such policies is to guarantee both a reasonable and predictable price to the consumer (assignative efficiency) through an efficient mechanism like that provided by an auction. In keeping with this line of thinking, Wolak (2000) claims that the introduction of fixed-price forward contract obligations increases competition and leads to supply prices that are closer to marginal costs. On the one hand, if suppliers raise their prices, they could end up selling less in the short-term market than in the long-term market. On the other hand, if the resulting market-

[1] Article published at Energy Policy Journal. Reference: Palacio, S.M., 2020. Predicting collusive patterns in a liberalized electricity market with mandatory auctions of forward contracts. Energy Policy Journal, 139, April 2020. DOI: https://doi.org/10.1016/j.enpol.2020.111311

clear price is high enough, it could generate higher opportunity costs than the gains to be made from exercising market power. Likewise, Woo et al. (2004) contend that electricity companies can reduce volatility and uncertainty by using forward contract purchases. Similarly, Strbac and Wolak (2017) explain that fixed-price forward contract obligations limit the incentive to exercise market power in the short-term market. Moreover, they argue that a large quantity of contracts of this type enhances competition, because if suppliers have enough quantity committed to fixed-price forward contract obligations, they will bid very aggressively to sell their output in the short-term market. Wolak (2017) presents empirical evidence of this effect in the Singapore electricity market, where the entrance of independent retailers competing with incumbents in the futures market yielded prices that were between 10-20% lower.

However, the success of these kinds of mechanisms is inconclusive and there are still several aspects that need to be studied, including the effect on daily markets. In this vein, our principal objective is to analyze the effect of the introduction of regulated long-term tariffs on liberalized electricity generation markets. Particularly, in this study, we examine the impact of mandated auctions on daily electricity markets and how collusion may impact daily prices. To do so, we seek answers to the following questions: Do regulated long-term tariff auctions trigger collusion in daily markets, i.e. will companies try to influence prices expectations in daily markets to get better deals in the auction market? Furthermore, to what extent could they be affecting daily prices? Here, we focus on the Spanish electricity market which provides us with a unique opportunity to analyze and use the *Contratos de Energía para el Suministro de Último Recurso* (CESUR) auction as a natural experiment.

The Spanish market is essentially divided between a daily and an intra-daily market, with the former accounting for most operations. Electricity generators offer the quantity of electricity they want to supply and the price at which they want to supply it for every hour of the day. Using a marginal price rule (lower price suppliers dispatch first), the market operator constructs demand and supply curves in real time, with the intersection being the equilibrium market price (corresponding basically to a uniform auction). In response to the problem outlined above, the CESUR auction emerged as a way to foster liquidity in long-term markets and to stabilize the consumers' tariff cost. Between 2007 and 2013, there were twenty-five CESUR auctions, which here serve as our natural experiment. In short, the auction was a long-term contract for a fixed quantity with the price being determined by a descending price auction. The auction ended when supply had satisfied demand. These CESUR auctions operated as a parallel market to the daily market and the contracts had a duration of between 3 and 6 months. The last CESUR auction was held on 19 December 2013. The next day, the Spanish energy regulator (National

Commission on Markets and Competition, CNMC) declared it invalid. The auction price was 7% higher than the daily price for the previous day. The CNMC explained that the fall in competitive pressure was a consequence of an unfavorable environment, based on low eolian production, high unavailability, a fall in trade on the inter-daily market, increasing demand, higher generation costs and a limited interconnection capacity. All these factors were essentially considered as being exogenous to the firms. Subsequently, no penalties were imposed. However, the CNMC may have failed to analyze the whole spectrum and perhaps the result was not so extraordinary after all, but rather a repeated hidden action.

Electric companies had several incentives to get higher CESUR prices. Besides risk premium, they received payments and discounts on the energy supplied in this market. The question is, how could they get better prices? There were two ways (see Fabra and Fabra Utray, 2012): (i) by taking off their supply offers during the auction (and, therefore, reducing the competitive pressure), and (ii) by affecting parallel market expectations, i.e, artificially increasing market daily prices the days previous to the auctions to get better deals on these. In this chapter we focus on (ii). We first present a theoretical framework which analyses two principal hypotheses: (a) the possibility that the inherent characteristics of these markets might trigger collusion and, as a result, "avoiding" pro-competitive regulation is a natural reaction; and, (b) that long-term tariffs in markets with excessive volatility may induce firms to try to reduce the adverse results by increasing their prices.

Second, and despite arguments in favor of the potentially positive effects of competition, we present empirical analysis which finds that the introduction of fixed-price forward contracts is associated with increases in electricity prices. A difference-in-difference-in-differences model estimates that the increase in prices was approximately 15% during the collusive phases. Moreover, ARMAX and LSTM simulations suggest that collusive agreements occurred 70 days before the CESUR auctions.

Following the above analysis, the main contribution of this chapter is to develop and validate the hypothesis of the existence of strong incentives to increase prices on daily markets before the regulatory policies of fixed long-term tariffs are applied. To the best of our knowledge, this is the first study that identifies that collusive behaviors in mandated auctions translate to daily markets, and that identifies and quantifies the collusive effect of mandated auctions over daily prices. This chapter also contributes to previous authors' analyses on CESUR auctions (Fabra and Fabra Utray, 2012; Capitan Herraiz and Monroy, 2014; Cartea and Villaplana, 2014; Peña and Rodriguez, 2018). In contrast to their papers, we focus on daily price increases instead of ex-post forward premiums. We first present an empirical model that detects abnormal price periods and then we compute the relative increase in daily

prices between normal and abnormal periods.

The empirical and the theoretical findings support our policy implications: The uncompetitive outcome seems to be explained by a highly concentrated market, a high elasticity to react to competitive regulations in a context of excessive volatility and the particularities of the auction design. Thus, well-designed mechanisms need to take into account the specific characteristics of the electricity market.

The rest of this chapter is organized as follows. In the next sections we present the literature review and our case study. In sections 4 and 5, we present the theoretical and the empirical models, respectively. In section 6, we present the results. Finally, in section 7, we conclude.

2.2 Literature Review

The process of deregulation in electricity markets where strong tendency to concentration exists, has introduced the necessity of tools that enhance competition. Forward contracts through auction mechanisms has been applied in many countries by regulators, based on the hypotheses that they can provide efficient production, competitive prices and foster investments.

Economic literature has traditionally argued that long-term forward contracts reduce market power. To illustrate this situation, a two Cournot duopolist model with N periods is presented in Allaz and Vila (1993). They show that, in equilibrium and with an increasing number of periods, duopolists will tend to the competitive solution if forward markets exist. In the same line, Green (1999) presents a model where generators with "Bertrand" conjectures in the forward market lead to the competitive solution in the spot market, and generators with "Cournot" conjectures do not participate in the contract market (unless they get a premium risk). The conclusion is that generators may hedge their output with forward contract sales. This will reduce the market power, because they will have a limited portion uncontracted in the spot market.

However, several authors claim that collusion is strongly related to auctions, particularly to repeated auctions (Graham and Marshall, 1987; McMillan, 1991; McAfee and McMillan, 1992; Porter and Zona, 1993; Aoyagi, 2003; Skrzypacz and Hopenhayn, 2004; Athey et al., 2004). Given that auctions exploit competition among agents, they create strong incentives to collude. As Green and Porter (1984) show, while collusion may keep profits higher than under no collusion, firms may learn to coordinate their strategies so as not to compete but still raise their discounted future benefits (Rotemberg and Saloner, 1986, and Haltiwanger and Harrington, 1991, draw similar conclusions). However, several authors argue that it is,

in fact, the auction format that will or will not trigger collusion. For example, Fabra (2003) shows that uniform auctions facilitate collusion more than do discriminatory auctions. Marshall and Marx (2009) find that cartels which control their members' bids can eliminate competition at second-price but not at first-price auctions. Pavlov (2008) and Che and Kim (2009) show how the information asymmetries in the auction design can be used to reach a competitive equilibrium. Benjamin (2011) finds that threats of future punishment allow players to reach a self-enforcing collusive equilibrium with greater payoffs than those of the static Nash equilibrium. Chassang and Ortner (2018) show that under collusion, bidding constraints can improve competition by limiting the scope for punishment.

In the field of energy economics, the theoretical analysis of auctions was initially developed by von der Fehr and Harbord (1992), Green and Newbery (1992) and von der Fehr and Harbord (1993), with particular reference to the structure of the UK market. Authors that include Fabra et al. (2002), Fabra (2003), Fabra (2006), Fabra et al. (2006), De Frutos and Fabra (2011), Fabra and Reguant (2014) and Fabra and Garcia (2015) likewise studied auction design in electricity markets, taking into account such factors as capacity, multiple offers, demand-elasticity, uncertainty and switching costs. On the empirical side, the seminal paper by Porter (1983) studies collusion in a railroad cartel that controlled eastbound freight shipments. The author uses a switching regression between periods of collusion and periods of competition, based on the stochastic process of being or not being in a collusive phase to evaluate his hypothesis. Other relevant studies that develop empirical methods to detect collusion are Ellison (1994) and Ishii (2008) who evaluate the empirical implications of the theoretical models of Green and Porter and Rotemberg and Saloner, by identifying price war patterns. Porter and Zona (1993, 1999) analyze bidding behavior in auctions for state highway construction contracts and a school milk procurement process, respectively. Borenstein and Shepard (1996) evaluate the conclusions of Haltiwanger and Harrington (1991). By using OLS and AR1 estimations, they find evidence of tacit collusion in the gasoline market in 60 cities between 1986 and 1992. Fabra and Toro (2005) model pool price patterns in Spain by means of an autoregressive Markov-switching model with time varying transition probabilities. Based on Bajari and Ye's (2003) approach to test for bid rigging in procurement auctions, Chassang and Ortner (2018) show that collusion is weakened by the introduction of bidding constraints in procurement data.

There is also a long tradition of the evaluation of price dynamics in electricity markets. Authors that include Engle (1982), Bollerslev (1986), Escribano et al. (2002), Goto and Karoly (2004), Leon and Rubia (2004), Worthington et al. (2005), Misiorek et al. (2006), Weron and Misiorek (2008) analyze price volatility using ARCH, ARX and GARCH processes. Cho et al. (1995), Huang (1997),

Huang and Shih (2003), Nowicka-Zagrajek and Weron (2002), Contreras et al. (2003), Cuaresma et al. (2004), Conejo et al. (2005), Zhou et al. (2006), Tan et al. (2010) and Yang et al. (2017) study price dynamics using ARIMA models. In the same line, several authors use ARIMA and SARIMA estimations to model hourly demand predictions (Ramanathan et al., 1997; Soares and Medeiros, 2005; Soares and Souza, 2006). Artificial neural networks and machine-learning models are also becoming quite popular for modeling prices: Szkuta et al. (1999), Fan et al. (2007), Catalao et al. (2007), Che and Wang (2010), Lin et al. (2010), Xiao et al. (2017), Wang et al. (2017). While there is plenty of academic literature about electricity price forecasting, we particularly recommend Aggarwal et al. (2009), Weron (2014), and Lago et al. (2018) who compare several machine-learning, deep-learning and linear models to forecast electricity spot prices.

Several studies have analyzed the case study of CESUR auctions. One of the first studies is Arnedillo Blanco (2011), who analyzes various concentration indices in the Spanish market during the period the auctions were held. Although he has found some statistical evidence that between 2009 and 2011 CESUR prices were systematically higher than the spot price, the evidence presented is inconclusive about market power over spot and CESUR prices. Fabra and Fabra Utray (2010, 2012) present and exhaustive analysis about market power and regulatory deficiencies in the Spanish market. They explain the perverse incentives that the CESUR auction introduced and that the main cause was a defective policy design. Although they provide very useful insights for our study, the analysis remains on basic statistical indicators. Peña and Rodriguez (2018) is probably the most detailed study on CESUR auctions.They analyze ex-post forward premiums and they find that winning bidders got a yearly average premium of 7.22%. What is interesting on this paper is that they are the first to analyze the full set of auctions, finding causal relations between number of bidders, spot price volatility and ex-post forward premium. Additionally, and supporting part of our main results, they find that hedging and speculative activities in derivative markets increases in dates near the auctions.

However, none of the reviewed studies considered modelling the impact of CESUR auctions on daily prices. We do include this aspect in our study by implementing a triple differences model-DDD- (Gruber, 1994; Berck and Villas-Boas, 2016). Aside from the advantages of using a double differences model (Angrist and Krueger, 1992; Card and Krueger, 1994; Meyer, 1995), we can additionally reduce the bias in our estimations by implementing a DDD model. As well, we identify the abnormal dynamic pattern of daily prices in dates near the auctions, by presenting an ARMAX model and a Long Short-Term Memory network. We will describe our empirical methodology in detail in Section 2.5.

2.3 Case Study

In 1997, the Spanish electricity market was liberalized (Law 54/1997). The market is divided between day-ahead and intraday operations, with the former concentrating most of the business. In the day-ahead market, agents submit the price and quantity offers that they are willing to accept over the next 24 hours. The market operator then constructs supply and demand curves by order of merit. The curve's intersection is the market clearing price (a marginal price system) that is paid to all suppliers offering a lower or equal market price.

The CESUR auctions were begun in 2007 in order to foster liquidity in spot markets and to stabilize price volatility. At these dynamic descending price auctions, agents competed in prices, and the winner was committed to supply energy by a fixed period.

In general, the CESUR auctions assigned contracts to supply users by means of a dynamic (descending) auction which was organized by rounds (the initial price being set by the market regulator). When the offers met the fixed demand, equilibrium was reached. Before the auction started, there was a phase in which qualified and pre-qualified agents were approved or not based on a set of normative requirements. Once approval had been granted, agents could then submit their offers. When the round was closed, the market operator analyzed the offers, if there was excess supply, a new round was opened with a lower price than the previous round's (see Figure 2.1).

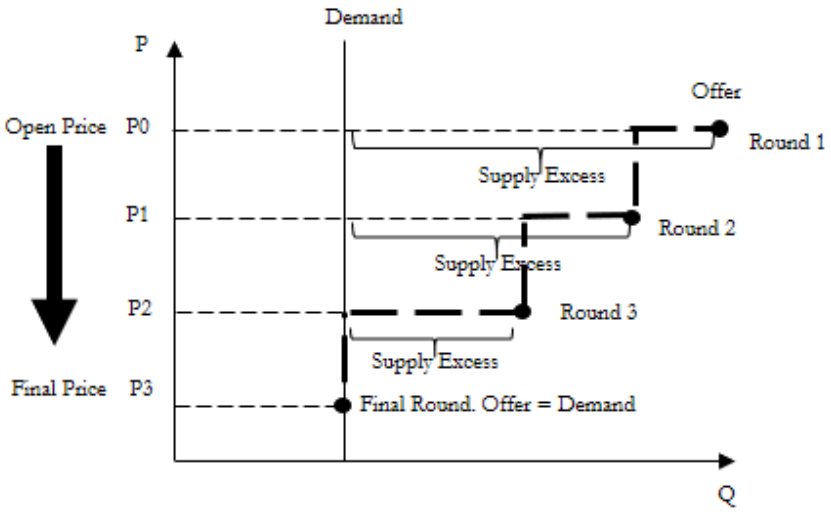


Figure 2.1. Auction Design

Under Law 54/1997, the organization, operation and liquidation of the CESUR auction were the responsibility of the Iberian market operator (OMIE). The su-

pervision and validation of the results was the job of the *Comisión Nacional de Mercados y la Competencia* (CNMC), as provided by complementary resolutions ITC/1659/2009 and ITC/1601/2010, and in keeping with principles of transparency, competitiveness and non-discrimination.

A total of twenty-five auctions were held. The last one took place on 19 December 2013, in line with the conditions laid down by the Energy Secretary's resolutions of 11 June 2010, 20 November 2013 and 11 December 2013 and with the criteria included in complementary resolutions ITC/1601/2010 and ITC/1659/2009. On 20 December 2013, however, the CNMC declared the auction void and published a report in which it offered the following reasons: Low eolian production, high unavailability, fall in transactions on the intraday market, increasing demand, rising generation costs and a limited interconnection capacity. According to the CNMC, these concerns led the auction participants to behave atypically. Qualified suppliers were lower in number and offers were withdrawn earlier than at other auctions. The auction was closed in the earliest round ever and, as a result, the resulting equilibrium price was 7% higher than that on the day-ahead market.

Date	**Auction Number**	**Qualified**	**Round Numbers**	**Awarded**	**AuctionAmount (MW)**	**Auction Price (€/MWh)**
15/12/2009	10	31	17	26	10,740	40.40
23/06/2010	11	33	14	30	4,536	45.21
21/09/2010	12	31	14	30	4,392	47.48
14/12/2010	13	25	12	22	4,306	49.42
22/03/2011	14	23	14	21	4,406	52.10
28/06/2011	15	26	17	23	4,288	53.75
27/09/2011	16	26	12	25	4,258	58.53
20/12/2011	17	28	19	28	4,363	53.40
21/03/2012	18	28	14	26	3,451	51.69
26/06/2012	19	29	18	25	3,575	57.09
25/09/2012	20	28	16	20	3,334	49.75
21/12/2012	21	28	18	30	3,345	54.90
20/03/2013	22	32	22	29	2,880	46.27
25/06/2013	23	34	18	48	3072	49.30
24/09/2013	24	37	12	44	2,852	48.74
19/12/2013	25	36	7	-	2,833	61.83

Table 2.1. CESUR Auctions. Columns contain date, the auction number, the number of qualified suppliers, the number of rounds, the number of winning bidders, the auctioned amount and the final auction price.

Table 2.1 shows that the number of qualifiers was similar to previous auctions, however, the final number of rounds was the lowest. In addition, the initial volume auctioned was lower than previous auctions (2,833 MW). During the first round, suppliers reduced their volume offers by 30.6% (the largest amount ever declined in a first round). Between 8 to 12 agents decided not to participate in the sec-

ond round. The second most important volume reduction was in the sixth round (15.2%), and the equilibrium was finally reached in the seventh round with a price of 61.83 €/MWh.

One of the characteristics of the CESUR auction was that the information about previous rounds was in aggregated terms, consequently, agents did not have precise and complete information about supply excess (i.e. they did not know if they were pivotal agents). However, in each round, they knew a range of supply excess. For example, in the 25th auction, during the first two rounds, agents knew that there was a supply excess of 200% (named the blind range). When the third round was finished, they knew that the supply excess was around 150-175% (4,250-4,958 MW offered versus 2,833 auctioned). As a result, qualified suppliers had information that the auction was outside the blind range (and, therefore, close to the equilibrium price). After the fifth round, they already knew that the supply excess was lower than 66%.

One of the most striking features of the decision of not applying penalizations was that the CNMC, despite finding evidence of abnormal prices and an unconventional auction, identified only exogenous factors as justification for their stance. Indeed, their understanding was that the low competitive pressure was caused by the negative environment affecting firms, and while they did not validate the results, they did justify their actions. Subsequently, no penalties were imposed.

The main reasons the CNMC argued were: a low eolian production, high unavailability, increasing demand, and higher generation costs. Figure 2.2 presents an overview of those factors. Top figure shows eolic source production over total sources production. During December 2013, 21% of total electricity was produced by eolic sources. The month previous to the 24th auction, eolic sources represented around 15% of total sources. Furthermore, the eolic production since the first auction was 14% on average.

Additionally, it is not visually clear in the second plot that a large increase in demand could explain this situation. During December 2013 16,699 GW were demanded, a very similar number to December 2012 (16,267 GW) while still far from other peak months like April (18,002 GW) or January (17,443 GW).

In terms of power unavailability, 4,961 GW were unavailable in December 2013. The average between July 2007 and November 2013 was 5,369 GW. Moreover, in the 24th CESUR auction, 8,386 GW were unavailable.

As fossil fuel sources are usually the most expensive technologies, we also show historical data about the International Petrol Price per barrel (USD) in Europe and the International Gas Price (USD/BTU). While the average petrol price during 2012 was 111.62 USD per barrel, in December 2013, the average price was 110.75, marginally higher than December 2012 (109.45) and lower than months like Oc-

tober 2013 and September 2013 (around 111 USD per barrel). The price of gas in December 2013 was slightly higher than the same month in 2012 (4.28 versus 3.45 USD/BTU), however, it was not excessively different from many other months during the period analysis.

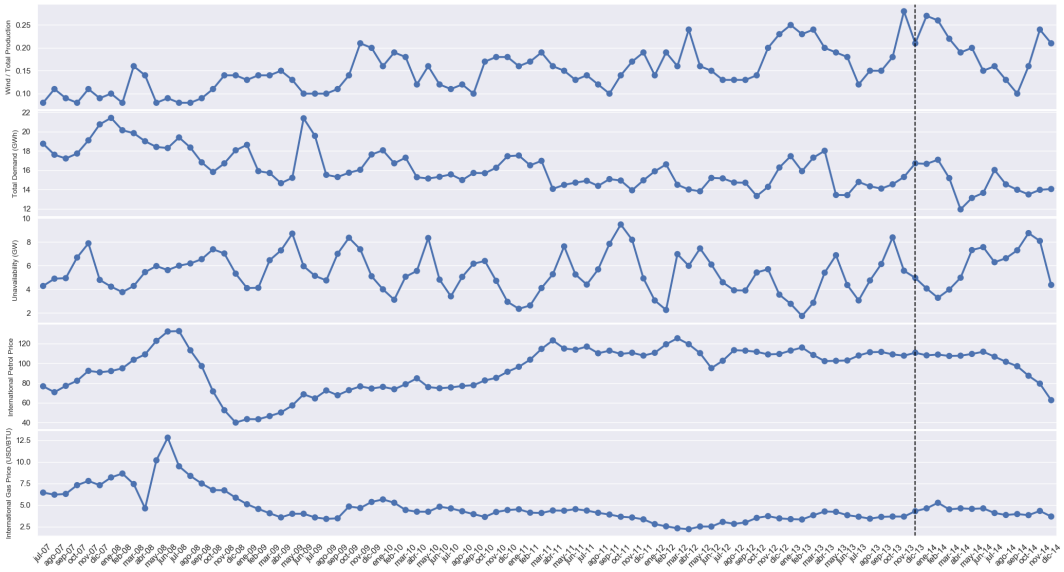


Figure 2.2. Main Factors Argued by the CNMC. Plots contain wind production over total production, total demand (GWh), unavailable power (GW), international petrol price (USD per barrel) and international gas price (USD/BTU).

On one hand, it is true that there was a lower volume auctioned and that there were higher declined volumes, on the other hand, it is hard to explain that there were exogenous factors that could explain this situation. However, it is still not clear if agents could extract information (due to the auction design) that may contribute to tacit collusion and if the output in this particular auction was an isolated case or was part of a repeated hidden action.

As was pointed out by Fabra and Fabra Utray (2012), Electric companies had several incentives to keep a higher CESUR price. Besides risk premiums to cover volatility in the spot market, there were also payments and discounts tied to the energy sell. The question is then: if collusion exists, which mechanisms through suppliers could have affected the CESUR price?

First, as we mentioned before, they could take off their bids, i.e. reduce competitive pressure. Second, they could alter expectations by artificially increasing daily market prices the days previous to the auctions. In this chapter, and as data about bidder identities or transactions on previous auctions is not public, we focus on the second mechanism.

2.3.1 Forward Premium[2]

The ex-post forward premium is defined as the difference between the auction price and the spot price during the delivery period. Competitive auctions imply ex-post forward premiums close to zero.

Table 2.2 compares prices for the auctions delivered by the market operator. The average ex-post forward premium during this period was 9.42%[3]. In only 3 out of 19 auctions, ex-post premium was negative. The same conclusion is stated by Peña and Rodriguez (2018). They do the same analysis considering the total number of CESUR auctions, finding an ex-post forward premium average of 7.22%. This translated into euros represented around 1,000 million overpaid cost for the consumers.

Date	Auction Number	Auction Price €/MWh	Daily Price (€/MWh)	Ex-Post Forward Premium (€/MWh)	Ex-Post Forward Premium (%)	Petroleum Price Variation (%)	Gas Price Variation (%)
25/09/2008	6	72.48	64.65	7.83	10.80 %	-52.2 %	-23.9 %
16/12/2008	7	56.47	43.1	13.37	23.68 %	-18.7 %	-29.1 %
26/03/2009	8	36.84	36.99	-0.15	-0.41 %	32.1 %	-15.7 %
25/06/2009	9	44.54	33.96	10.58	23.76 %	20.6 %	3.4 %
15/12/2009	10	40.40	39.96	0.44	1.10 %	8.7 %	21.9 %
23/06/2010	11	45.21	44.07	1.14	2.51 %	-0.2 %	-11.8 %
21/09/2010	12	47.48	43.33	4.15	8.74 %	9.3 %	-8.7 %
14/12/2010	13	49.42	45.22	4.20	8.50 %	21.0 %	9.5 %
22/03/2011	14	52.10	48.12	3.98	7.64 %	15.5 %	3.1 %
28/06/2011	15	53.75	54.23	-0.48	-0.89 %	-3.4 %	-6.7 %
27/09/2011	16	58.53	52.01	6.52	11.14 %	-3.5 %	-14.7 %
20/12/2011	17	53.40	50.64	2.76	5.17 %	8.3 %	-27.7 %
21/03/2012	18	51.69	46.07	5.62	10.87 %	-8.6 %	-5.3 %
26/06/2012	19	57.09	49.09	8.00	14.02 %	1.2 %	22.5 %
25/09/2012	20	49.75	43.16	6.59	13.25 %	0.4 %	20.8 %
21/12/2012	21	54.90	40.34	14.56	26.52 %	2.2 %	-1.6 %
20/03/2013	22	46.27	34.26	12.01	25.96 %	-8.8 %	16.9 %
25/06/2013	23	49.30	49.81	-0.51	-1.03 %	7.5 %	-13.0 %
24/09/2013	24	48.74	54.73	-5.99	-12.28 %	-1.7 %	2.8 %

Table 2.2. Auction Prices Compared to Daily Prices

A reason usually argued for why long-term contracts auctions may result in positive ex-post forward premiums is that bidders are expecting a risk premium. As they are taking risk by selling long-term fixed contracts, they should be able to cover their generation costs. This is mainly relevant for fossil generators that are tied to international price fluctuations. However, as we will see later, Spain has quite a diversified energy source structure where fossil fuels are not the main resource. In addition, columns in Table 2.2 contain petroleum and gas price variations during delivery periods. If the risk premium argument would be true, negative price variations should be related to non-positive ex-post forward premiums, and vice versa, although a majority of periods with negative variations have had positive premiums.

[2]We thank an anonymous referee for this suggestion.

[3]The ex-post forward premium is calculated as $\frac{P_{CESUR}-P_{Spot}}{P_{Spot}}$

2.3.2 Sector Concentration

Since 1997, privatization and liberalization processes in Spain's electricity generation market have tended to concentrate the sector. At the start of the auctions, two enterprises accounted for 64% of the country's generation capacity (Agosti et al., 2007). Table 2.3 summarizes the shares of net power installed [4].

Firm	Net Power (MW)	Shares
Iberdrola Generación S.A.	20,017	34.8%
Endesa Generación S.A.	16,614	28.9%
Unión Fenosa Generación S.A.	5,959	10.4%
Gas Natural SDG, S.A.	2,791	4.9%
Hidroeléctrica del Cantábrico, S.A.	2,428	4.2%
Enel Viesgo Generación, S.L.	2,259	3.9%
Others	7,408	12.9%
Total MW	57,476	

Table 2.3. Installed Capacity by firms

Source: Agosti et al. (2007)

These two firms had a pivotal index rate[5] below 110% for more than 5% of the time, the threshold for considering the existence of a power market according to the European Commission. Moreover, interconnection with Europe was limited (excluding Portugal and Morocco). More particularly, Spain was under the 10% interconnection threshold recommended by the European Commission. Could these factors have contributed to price collusion?

It is therefore not surprising that after the 25th auction, 80% of the total auctioned volume was distributed between two firms: Iberdrola and Endesa[6] (see Table 2.4).

In addition to the inherent characteristics of the Spanish market, other factors could have affected prices. For instance, the external volatility may have adversely affected the firms' price decisions. According to this hypothesis, the design of the auctions (a fixed price during a fixed period) incentivized firms to charge a risk bonus which was also transferred to the daily market.

Determining whether prices are driven by external shocks is important for any subsequent impact analysis. Here, we propose exploiting the introduction of the CESUR auction across time and the market to identify the causal effect of introducing fixed long-term tariffs on prices using a triple difference-in-difference approach.

[4] This excludes generators with a power below 50 MW.

[5] The pivotal index rate seeks to show whether it is possible to supply the prevalent demand without a particular supplier.

[6] Union Fenosa was acquired by Gas Natural in July 2008.

Auction Price (€/MWh)	61.83
Auction Amount (MW)	2,833
Iberdrola Generación S.A.	924
Endesa Generación S.A.	1,336
Gas Natural SDG, S.A.	480
Others	93

Table 2.4. Winning Bidders - Auction 25th

2.3.3 Nord Pool

Nord Pool AS is the electrical energy market operating in Norway, Denmark, Sweden, Finland, Estonia, Latvia, Lithuania, Germany and the UK (but, note, that during the period analyzed Germany and the UK were not yet members and Lithuania did not join until June 2013). Here, we use the day-ahead Nord Pool market as a control group for the Spanish market given that the two markets were not related during the period analyzed.

As mentioned above, the CESUR auction was operational between 2007 and 2013. However, we should highlight that data for the control group are only available from 2013 onwards. Therefore, our estimations here refer to the period of days between 2013 and 2014, where 2013 is considered as a treated year (during which three auctions were held) and 2014 as a control year.

Table 2.5 shows total production (MWh) by country in the period analyzed. As can be seen, Norway and Sweden produced more than 70% of Nord Pool's total production.

Year	Norway	Sweden	Finland	Denmark	Latvia	Lithuania	Spain
2013	133,385,250	147,770,389	65,952,798	32,491,906	2,795,241	3,543,888	186,569,806
2014	141,158,884	149,710,633	64,587,200	30,648,162	4,903,420	2,959,173	170,399,215
Shares	35%	38%	17%	8%	1%	1%	-

Table 2.5. Total Production (MWh) by Country

Table 2.6 reports the sources of electricity generation by country. While Spain has quite a homogeneous structure, Nord Pool presents, overall, a highly dependent generation structure. For instance, 97% of Norwegian and 48% of Swedish production are extremely dependent on hydraulic sources.

2.4 Theoretical Model

We construct a very simple model to illustrate the possible implications of introducing fixed-price contract obligations. We consider two symmetric firms which sell

Source	Norway	Sweden	Finland	Denmark	Spain
Eolic	1.08%	4.40%	0.74%	26.28%	21.10%
Other Renewables	0.00%	6.69%	15.36%	31.89%	20.00%
Fossil Fuels	2.30%	2.85%	25.26%	41.84%	23.60%
Nuclear	0.00%	37.92%	32.64%	0.00%	20.40%
Hydraulic	96.62%	48.14%	24.67%	0.00%	11.80%
Others	0.00%	0.00%	1.33%	0.00%	3.10%

Table 2.6. Share of Electricity Generation by Country

a homogeneous good with constant marginal cost c. We also assume that they are risk neutral. The firms offer q which is covered perfectly by the demand.

We consider two auction formats: a uniform-price auction (the price received is the market price) and a discriminatory-price auction (the price received is equal to its own bid).

We have two markets: a daily market where the price mechanism is a uniform-price auction and a discriminatory-price auction which works every $2t$ periods. The timing of the game is represented as follows: Each firm simultaneously and independently submits a bid in the daily market specifying the minimum price at which it is willing to supply. As demand perfectly matched supply, both firms dispatch their total production. On the other market, the firms submit different bids b_i^s, and the lower bidder is the only firm that can deliver at that price. However, the quantity that is produced in this market (a fixed quantity $\bar{q}$ set by the auctioneer) cannot be delivered in the daily market.

Formally, the quantity produced by the firm i ($i = 1, 2$) in the daily market is given by:

$$q_i^d = \begin{cases} q_i - \bar{q} & b_i^s \leq b_j^s \\ q_i & b_i^s > b_j^s \end{cases} \tag{2.1}$$

The lowest accepted bid b^d in the auction is referred to as the market price P, and can be expressed as:

$$P = \begin{cases} b_j^d(q_j) & b_j^d \leq b_i^d \\ b_i^d(q_i) & b_j^d > b_i^d \end{cases} \tag{2.2}$$

In the secondary market (the discriminatory-price auction) the price is set by the lowest auction, i.e., $b^s = min(b_i^s, b_j^s)$, and the quantity produced is:

$$q_i^s = \begin{cases} \bar{q} & b_i^s \leq b_j^s \\ 0 & b_i^s > b_j^s \end{cases} \tag{2.3}$$

At the end of each stage, the two firms receive their profits. The auctioneer announces the market price in the daily market and which firm is to deliver in the secondary market (if it is a $2t$ period).

We explore an infinitely repeated game, with a strategy profile (S_i, S_j) and the payoff for each firm is the sum of their discounted profits, where $\rho \in (0,1)$ is the discount rate.

If the firms collude they will set a price $\bar{P}$ in the daily market and a price $\bar{b}^s$ in the secondary market. However, as the winning firm in the secondary market is not able to sell $\bar{q}$ in the daily market, it should earn at least $\bar{b}^s \geq \bar{P}$. The discounted future benefits if the firms i collude are:

$$\pi_i^C = (\bar{P} - c)q_i^d(1 + \rho + \rho^2 \dots) + (\bar{b}^s - c)q_i^s(\rho^2 + \rho^4 + \dots)$$

When the firms do not collude, they compete in the daily market and set a competitive price $\underline{b}$ in the secondary market. However, the firms set a $\underline{b}$ bid equal at least to the daily market price (i.e., $\underline{b} \geq P$). If the firms i do not collude, future discounted benefits can be expressed as:

$$\pi_i^{NC} = (P - c)q_i^d(1 + \rho + \rho^2 \dots) + (\underline{b} - c)q_i^s(\rho^2 + \rho^4 \dots)$$

We consider trigger strategies where the firms sustain a collusion price in the set $\bar{P}, \bar{b}$ in each period if and only if the firms do not deviate in previous periods, i.e., the collusion price path is an equilibrium of the perfect subgame equilibrium if and only if:

$$\pi^C(t, \rho) \geq \pi^{NC}(t, \rho) \forall t$$

Formally,

$$\pi^C - \pi^{NC} = \frac{1}{1 - \rho}(\bar{P} - P)q_i^d + \sum_{t=1}^{\infty} \rho^{2t}[\bar{b} - \underline{b}] \geq 0$$

The left side of the equation corresponds to the benefits of colluding in the daily market, while the right side corresponds to the benefits of colluding in the secondary market. If the firms do not collude in the secondary market, they can deviate in the secondary market and set a bid $\bar{b} - \epsilon$ and obtain all the profits. This means that firms will compete until they obtain a price P, which guarantees them at least the same profits as in the daily market. Therefore, the low boundary bid has to be $\underline{b} \geq P$. Likewise, for both firms to make a profit, and not to deviate in the secondary market, one firm has to profit from losing the secondary market auction, i.e., obtain profits in the daily market, that is, $\bar{P} \geq P$.

Proposition 1 *Exists $\hat{\rho} \in (0,1)$ such that the set price $\{\bar{b_i}, \underline{b_j}, \bar{P}, P\}_{t=1}^{\infty}$ is a sustainable equilibrium of the perfect subgame if and only if $\rho \in (\hat{\rho}, 1)$.*

What is interesting about the results is that the fixed-price contract (the secondary market auction) introduces an additional effect in the collusive pattern. Without this, the incentives to collude would be the same as those not to collude, because there would be no motivation to raise prices in the daily market. Yet, with the introduction of the secondary market, both firms have to set higher prices to ensure that the auction loser obtains at least the same benefits as those obtained by the winner. To guarantee that the loser obtains the same benefits for the $\bar{q}$ ratio, the firms have to set a price $\bar{P}$ at least higher than that of the bids (b_i^s, b_j^s).

Proposition 2 *Consider ϵ, an exogenous positive price shock which affects the daily market price immediately after the second market auction is finished. Then, the subgame equilibrium is one where $b > b^*$, where b^* is the price obtained in the secondary market when there are no shock prices.*

To prove Proposition 2, we assume an $\epsilon > 0$ which affects spot market prices immediately after the second market auction is held, with probability p. To simplify, we assume one of the firms always loses the auction and the other always wins. We can express the future discounted benefits of the firm that wins the auctions as:

$$\pi_W = (P-c)(q_W - \bar{q})(1+\rho+\rho^2\dots) + (b-c)\bar{q}(\rho^2+\rho^4\dots) + p\epsilon(\rho^2+\rho^4\dots)(q_W-\bar{q})$$

And the future discounted benefits of the firm that loses as:

$$\pi_L = (P-c)q_L(1+\rho+\rho^2\dots) + p\epsilon(\rho^2+\rho^4\dots)q_L$$

For colluding firms, the price path is an equilibrium of the perfect subgame equilibrium if and only if:

$$\pi_W(t,\rho) = \pi_L(t,\rho) \quad \forall t$$

Then, we can rearrange as:

$$\pi_W - \pi_L = (P-c)(q_W - q_L)\frac{1}{1-\rho} + p\epsilon \sum_{t=1}^{\infty} \rho^{2t}(q_W - q_L) - \bar{q}\left[(P-c)\frac{1}{1-\rho} + p\epsilon \sum_{t=1}^{\infty} \delta^{2t}\right] + \sum_{t=1}^{\infty} \rho^{2t}(b-c) = 0$$

To simplify the results (which does not affect the general conclusions), we assume that both firms produce the same quantity ($q_W = q_L$):

$$-(P-c)\bar{q}\frac{1}{1-\rho} - p\epsilon \sum_{t=1}^{2t} \bar{q} + \sum_{t=1}^{\infty} \rho^{2t}(b-c)\bar{q} = 0$$

The first left-hand term is the opportunity cost of the winning firm not selling $\bar{q}$ in the daily market. The second term is the opportunity cost of its not deriving the positive shock in the quantities sold $\bar{q}$. Finally, the third term is the profit made from winning the auction in the secondary market. Without a shock, the second term is null, and therefore, the profits of the secondary market auction derived from Proposition 1 have only to cover the opportunity cost of not selling in the daily market. However, in this new equilibrium, to be sustainable (with ρ, c and $\bar{q}$ fixed), i.e. to keep this equality, b must be higher than b^*. Also P is higher than the equilibrium price without a positive shock.

A number of preliminary conclusions can be derived from these two propositions. Artificially applying fixed-price auctions of this kind creates incentives to increase prices in the daily market. As seen, this effect could emerge for two reasons: First, repeated auctions trigger collusion because of their inherent format and because of the structural collusion present in the daily market (Proposition 1). Second, uncertainty in the daily market pushes firms to follow the expected price path in the secondary market (Proposition 2). In conclusion, for whatever reason, what triggers collusion is not the introduction of the regulation per se, but the actual design of that mechanism. In the following section, we present the empirical methodology to measure the size of this effect.

2.5 Methodology

In the following sections we describe our methodology to evaluate the impact of introducing fixed long-term tariffs on prices in the Spanish case. We are interested in two aspects: a) identifying abnormal patterns in the daily market due to the introduction of the CESUR auction, b) measuring the economic impact in terms of

higher daily prices. As we explained previously, electric companies had incentives to get higher CESUR prices. The mechanism we are focusing on is one in which firms could be influencing parallel markets to inflate CESUR price expectations. Therefore, we divide our analysis as follows: in Section 2.5.1, we present an ARMAX model combined with Instrumental Variables (IV) to identify if there were abnormal behaviors on daily prices during the previous days to the auctions, i.e. find evidence that firms were inflating expectations in the daily market to get better prices in the CESUR auction. Once we identify abnormal price behavior we propose in 2.5.2 a triple differences model (DDD) to measure the economic impact on energy prices.

2.5.1 Price Dynamic in Electrical Markets

The dependent variable in our analysis is the logarithm of the daily price. The database includes daily data on prices and production obtained from OMIE and Nord Pool during the period 2013-2014. The weather condition variables are from National Centers for Environmental Information. We also include the International Petrol Price in Europe, the International Price of Gas (USD/BTU) and Spain's Risk Premium relative to Germany. Finally, three auctions were held in 2013: 20 March 2013, 25 June 2013 and 24 December 2013.

If there were abnormal price phases during the period analyzed (whether resulting from collusion or not), the question is how these periods can be best defined. Our main objective is to understand when the dynamics of prices in the Spanish market and that of prices in the Nord Pool market present a different pattern. To do so, we add a dummy variable (IP - Investigation Period) before, during and after the auctions that is equal to one for each specific time window. We test our results for the three auctions held in 2013. In addition, given that the auctions were canceled in 2014, we can use this year as a falsification test.

As we are modeling electricity prices that are simultaneously set by supply and demand, several questions have to be addressed before we continue. Ordinary least square (OLS) estimates are based on the assumption that all independent variables are uncorrelated with the error term (exogeneity assumption). When this occurs, the conditional expected error term is equal to zero for each observation, which implies that the regressors are orthogonal to the error term.

However, we might expect the error term to be correlated with quantity, given that quantity depends on other factors (omitted bias). Additionally, the prices and quantities observed reflect equilibrium sets between supply and demand, i.e. they are determined simultaneously and the estimation may therefore be affected by the simultaneity bias (Angrist and Krueger, 2001). As such, we cannot infer if changes

in price are due to supply or demand shifts.

In addition, the nature of the product (perfect synchronization without storage capacity) causes high price volatility. This means the price is highly affected by its dynamics which inherits some of the distinctive features of demand, such as seasonality effects (Fezzi, 2007). Moreover, supply and demand must be constantly balanced, but as shocks cannot be smoothed, they are immediately transferred to prices.

All these issues are crucial elements to take into account when estimating electricity price dynamics: Simultaneity bias, non-stationarity, non-normality and serial correlation should each be carefully examined.

Formally, we initially define an autoregressive moving average model with exogenous input terms (ARMAX) that can be expressed as:

$$y_t = \beta X_t + \gamma IP_t + \alpha_1 Q_t^{IVSpain} + \alpha_2 Q_t^{IVNPool} + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{m} \theta_j \mu_{t-j} + \mu_t \quad (2.4)$$

where y_t is the dependent variable and is equal to $log(P_{Spain}/P_{NPool})$. P_{Spain} is the daily arithmetic average price of the spot market in Spain, and P_{NPool} is the daily arithmetic average price of the spot market in Nord Pool. IP is the indicator which reflects switching phases and Q_t^{IV} is the quantity production of each market (Spain, Nord Pool expressed in logarithms), both quantities being instrumental variables (see section Appendix A: Instrumental Variables Tests). X is the vector that represents explanatory variables and the model orders p, m refer to p autoregressive and m moving average terms.

Following the practical approach developed by Box et al. (2015) to select an appropriate ARMA model with the corresponding extensions applied to ARMAX models (Andrews et al., 2013) – see Appendix B: ARMAX Model-Building Process – we choose our main exogenous candidate variables which are reported in Table 2.7. All the variables are expressed in MWh (as logarithms). Dummy Null Price are two atypical days in April 2013, when prices in Spain were equal to zero.

2.5.2 Identification and Estimation Methods

Our second objective is to identify the average effect of mandated auctions on prices in the market and the year in which regulations of this type were introduced. Specifically, we are interested in comparing prices the days before mandated auctions were introduced with the counterfactual, i.e. prices in a different market on the

Exogeneous-Variables Candidates
Nord Pool Production
Spain Production
Portugal Production
Spain Production Lag-1
Spain Imports
Dummy Null Price

Table 2.7. Stepwise Regression Variable Candidates

same days.

A major concern is that the country that chose to introduce the regulation is likely to be different from the country that did not and that such differences may be correlated with prices. In principle, many of the unobservable characteristics that may confound identification are those that vary across markets but are fixed over time. A widely used method for controlling time-invariant unobserved characteristics and unexpected variations is to use panel data and estimate difference-in-differences models.

A difference-in-differences estimation (Angrist and Krueger, 1992; Card and Krueger, 1994; Meyer, 1995 are the precursors) measures the impact of a policy change on an outcome variable by removing the effects of time and place. When the outcome depends on the policy, time, place or other variables, we can additionally reduce the bias in the estimated effect by using a triple difference-in-differences (Berck and Villas-Boas, 2016). The goal of this empirical study is to capture the effect of fixed price long-term auctions held in 2013 (experimental year) and which impacted Spain (treatment group). Identifying this effect implies controlling at the same time for any systemic shocks that could have disturbed the electricity market outcomes of the treatment group in the experimental year. Higher prices in Spain when the regulations are in place can then be considered as the net effect of regulation. To control for economic changes that are unrelated to the program, we use a market (Nord Pool) which was unaffected by the regulation.

The overall difference-in-differences can be written as:

$$DD = E[(Y_t^T - Y_{t-1}^T)|IP_t = 1] - E[(Y_t^C - Y_{t-1}^C)|IP_t = 0]$$

where Y_t^T is the treated group in time t (Spain, 70 days before each auction in 2013), Y_t^C is the control group in time t (Nord Pool, 70 days before each auction in 2013) and IP_t is an indicator of the policy effect.

Recall that, in 2014, the law governing CESUR auctions was lifted; thus, we can use 2014 as an additional control by using triple differences (DDD):

$$DDD = E[(Y_t^T - Y_{t-1}^T)|IP_t = 1] - E[(Y_t^C - Y_{t-1}^C)|IP_t = 0] \\ - E[(Y_{t+2}^T - Y_{t+1}^T)|IP_{t+2} = 1] - E[(Y_{t+2}^C - Y_{t+1}^C)|IP_{t+2} = 0]$$

where Y_{t+2}^T is the treated group in time $t+2$ (Spain, 70 days before the same day the auctions where celebrated but in 2014), and Y_{t+2}^C is the control group in time $t+2$ (Nord Pool, 70 days before the same day the auctions were held but in 2014).

The left part of the equation expresses two concepts: first, the difference between daily prices of the treatment and control group 70 before each auction ($Y_t^T - Y_t^C$). This term should be positive if mandated auctions introduce incentives to increase prices. Second, the difference between daily prices of the treatment and control group during previous periods considered as untreated ($Y_{t-1}^T - Y_{t-1}^C$). This control for unobservables that affect differences in daily prices and are constant over time.

The right-part of the equation introduces another two concepts: first, the difference between treatment and control group in the non-experimental year ($Y_{t+2}^T - Y_{t+2}^C$).It controls market individual differences that may affect prices in days in which auctions would have been applied (in 2014) if they were not cancelled. Second, ($Y_{t+1}^T - Y_{t+1}^C$) controls for market individual differences that affect prices during the same days that auctions would have not been applied. Overall, it controls omitted variables that cause differences in daily prices to change over the period.

This can be simplified as:

$$DDD = DD_{2013} - DD_{2014}$$

DDD will consistently identify the effect of the policy if two conditions hold: First, the differences in prices in 2013 are related to the auctions, i.e., $DD_{2013} > 0$. Second, there are no differences in prices in the counterfactual year, i.e., $DD_{2014} = 0$.

2.6 Results

In this section, the model results are analyzed. As mentioned before, we first identify abnormal periods in the daily market. Second, we present our identification strategy and, third, we measure the average effect of mandated auctions on prices.

2.6.1 Windows Choice

We present the equation (1) estimation results in Table 2.8. We analyze whether daily prices are consistent with the hypothesis of expected inflation near auction dates. Our focus is on the Investigation Period variable (IP), which is a dummy that takes one during days surrounding auctions. We test whether there are significant differences between Spanish daily prices and Nord Pool daily prices. Additionally, we validate our analysis conducting the same experiment during a year without auctions.

The table highlights that, during 2013, there are significant, positive and persistent price differences between 0 and 70 days before the CESUR auctions. The differences are between the daily average Spanish prices and the daily average of the Nord Pool market. Those differences oscillate between 68% and 222%. After that, we find no conclusive evidence of increasing prices.

We also report estimations during 2014 of the Investigation Period variable. As we mentioned earlier, auctions were cancelled after December 2013. Therefore, we would expect no relation between the Investigation Period variable and daily prices. The results from the simulation in 2014 are also displayed in Table 2.8. As expected, we do not find any evidence of price differences, which initially supports the hypothesis that regulation negatively affected consumers.

Days	**-90**	**-85**	**-80**	**-75**	**-70**	**-65**	**-60**	**-55**	**-50**	**-45**	**-30**	**-25**	**-20**	**-15**
IP 2013	-0.07	0.31	0.35	0.37	0.84***	0.83***	0.52**	0.72***	0.65**	1.17***	0.85***	0.62**	0.63***	0.86***
IP 2014	0.20	-0.18	-0.05	0.10	-0.05	0.08	-0.02	0.07	0.32	-0.19	-0.08	-0.04	-0.05	0.02

Days	**-10**	**-7**	**-5**	**-3**	**0**	**3**	**5**	**7**	**10**	**15**	**20**	**25**	**30**	**45**
IP 2013	0.34	0.54**	0.92***	0.70***	0.79***	-0.99***	-0.41	1.11***	0.56*	0.71**	0.14	-0.34	-0.33	-0.15
IP 2014	0.22	-0.05	-0.03	0.22	-0.02	0.03	0.06	0.23	-0.01	-0.20	0.09	0.12	0.06	0.28*

Table 2.8. Window Time Choice

To validate that 70 days before each auction is representative of abnormal behavior in daily prices, we propose a long short-term memory network model (see Appendix C for a complete specification of the model). Here, our main concern is to demonstrate that the abnormal period of collusion chosen is in fact correct. To do so, we employ a deep learning model operating as an outlier detection model. The advantage of models of this type is that they are well suited to predicting patterns. Although we cannot assess the marginal impact of the variables, we can use these models to verify if our switching regression pivotal variable is correctly defined.

In Figure 2.3 we illustrate the results for the normal observations. The upper panel of the figure plots the true values, predicted values and error values, while the lower panel shows the logarithm of the probability density function of the error test using the mean and the covariance of the training error vector. If values are below the "normal" threshold, they are considered abnormalities.

Figure 2.3. Normal Data Test

In this particular case, when testing the normal data for 2013, only five days were considered to be abnormal. However, when we focus on the 70 day-period before the CESUR auctions (see Figure 2.4), almost every day is abnormal. Specifically, we obtain a Precision of 100% and a Recall of 81%[7].

Figure 2.4. Abnormal Data Test

Therefore, empirical evidence supports the hypothesis of inflating prices the day previous to the auction. The introduction of long-term tariff auctions is related to large price differences between Spanish and Nord Pool markets. Thus, our choice

[7]Precision = True Negatives / (True Negatives + False Positives), Recall = True Positives / (True Positives + False Negatives)

of the 70-day time period prior to the CESUR auctions as our main candidate seems to be accurate. Indeed, during this period, we captured recurrent abnormal patterns in energy prices.Hereinafter, therefore, we propose using the dummy 70 days before a CESUR auction as our collusive or abnormal phase variable. We now turn to the question of what the net impact on prices when long-term tariffs were introduced. To shed light on this point, we compute triple differences between Spanish and Nord Pool markets and between 2013 and 2014.

2.6.2 Identification Strategy

Although market regulators introduce mandated auctions to foster competition and provide energy at a competitive price, we have found that the particular characteristics of electricity markets (high concentration and high volatility) may produce undesired effects on daily markets. In this section, we study to what extent daily prices were affected by the introduction of regulation.

To measure the effect on the actual price increase due to the CESUR auctions, it would be necessary to observe how the electric market behaves in two different states: one in which the regulation was applied, and one in which it was never applied. The increase in prices when the auction is held compared to when it is not would therefore be the actual effect of the regulation. As it is impossible to observe the same market in a counterfactual state, the best approach is to use "comparables". Thus, it is necessary to create counterfactuals that are nearly identical and which will enable us to control for additional sources of omitted variable bias.

However, it is almost impossible to have two virtually equal markets. Here is where a triple differences estimation is useful. A difference-in-differences model controls for omitted variables that could affect differences in daily prices between the Spanish market and the Nord Pool market, and that are constant over time. The main advantage of using a DDD model is that, in addition to controlling for those factors, we will also be able to control for omitted variables that affect differences in daily prices over time for each of those markets. This will help us get better unbiased estimates of the true effect of mandated auctions.

Table 2.9 summarizes individuals' statistics in the treatment and the control group and in experimental and non-experimental years. The Difference column shows difference in means tests. In Spain, there were almost no significant differences in the source production structure. However, there is a decrease in the fossil production (explained by a decrease in combined cycle power production) that can be related to oil and gas price variations (which are also significant). Therefore, it is important to include them as control variables. Eolic production also has a significant difference, but we believe that this difference is quite small and should not bias the results.

Spain	2014	2013	Difference
Portugal Production (MW)	133,168	132,258	-909.6
	(28,650.9)	(24,185.1)	(24,518.5)
Unavailability (MW)	199,384	155,749	-43,635***
	(60,790.5)	(64,542.2)	(48,173.5)
Hydraulic (MW)	56,434	52,555	-4878,5
	(32,320.1)	(34,864)	(37,143)
Nuclear (MW)	37,306	34,542	-2,764.6
	(28,634.1)	(28,042.1)	(40,547.4)
Fossil (MW)	299,025	334,845	35,820***
	(64,232)	(67,410)	(79,868)
Eolic (%)	0.19	0.20	0.007**
	(0.05)	(0.04)	(0.02)
Oil Price (USD/Barrel)	99.05	108.62	9.57***
	(14.76)	(4.63)	(16.38)
Gas Price (USD/BTU)	4.31	3.76	-0.50***
	(0.49)	(0.36)	(0.65)
Risk Premium	149.69	144.57	-5.12
	(27.76)	(48.69)	(28.32)
Temperature (C^o)	16.83	17.16	-0.3
	(5.48)	(6.08)	(2.88)
Precipitations (mm)	1.42	1.63	0.210
	(6.85)	(5.62)	(9.07)
Workday=1	0.70	0.69	-0.01
	(0.46)	(0.45)	(0.11)
Nord Pool	**2014**	**2013**	**Difference**
Quantity (MW)	1,060,728	1,040,001	-20,981**
	(186,611)	(197,833)	(86,992)
Precipitations(mm)	0.06	0.07	0.008
	(0.12)	(0.14)	(0.179)
Temperature (C^o)	1.11	1.07	-0.04
	(0.47)	(0.54)	(0.38)
Workday=1	0.71	0.70	-0.008
	(0.46)	(0.46)	(0.53)

Table 2.9. Difference in Means between Experimental-Non Experimental Years and Control-Treatment Groups

There is quite an increase of unavailability power during 2014. This increase should have put more pressure on prices in 2014 and, therefore could have hidden the auction effects. Then risk premium, Portugal production, and weather conditions do not have significant differences between experimental and non-experimental years. In terms of production, the Nord Pool market has a significant increase in 2014, however this is related to the entry of Lithuania and, as a consequence, we will control for this factor. We do not find evidence of different weather conditions between experimental and non-experimental years. In summary, we control for fossil prices, unavailability, and the Lithuanian entry. Due to a limited number of observable characteristics, there are potential omitted variables that may affect the price in each group and year differently. However, as we are using data from across markets and time periods, we expect that with a triple differences model we can control for unobservable variables that affect prices across groups and over time.

The identifying assumption of the DDD (Gruber, 1994) only requires that there be no shock that affects the relative outcomes of the treatment group in the same country-year as the regulation.

In short, our treatment group is Spain while our control group is Nord Pool (i.e., the market unaffected by the policy). However, the counterfactuals are also considered between experimental and non-experimental years.

Therefore, we transform our data-set to panel data and adopt a triple difference-in-differences approach. By comparing changes, we control for observed and unobserved time-invariant market characteristics that might be correlated with the introduction of the policy decision as well as with prices.

Formally, the DDD model can be specified as:

$$P_{ijt} = \beta_1 x_{ijt} + \beta_2 IP_t + \beta_3 \rho_j + \beta_4 T_i + \beta_5 \rho_j IP_t + \beta_6 IP_t T_i + \beta_7 \rho_j T_i + \beta_8 \rho_j IP_t T_i + \epsilon_{ijt} \quad (2.5)$$

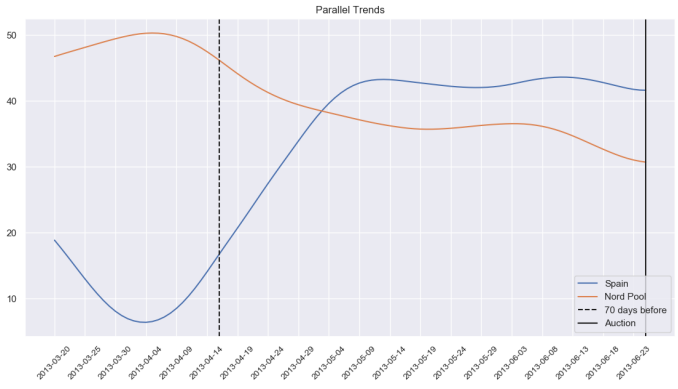
where P_{ijt} is the logarithm of the price in the market i (Spain, Nord Pool), in the experimental year j (2013, 2014) on day t. IP_t is an indicator variable that takes a value of one 70 days before the auctions and zero otherwise. T_i is a fixed effect unique to market i and ρ_j is a time effect common to both markets in year j. x_{ijt} is a vector of covariates that vary across markets and time. Among the control variables, we maintain the autoregressive terms and the instrumental variables to control for the biases discussed above (see Bertrand et al., 2004, for a detailed discussion).

β_1 are several control variables that are unique for each market and causes differences in daily prices (We include Oil Price, Gas Price, Power Unavailability, Autoregressive Terms, a production variable estimated through instrumented variables, a dummy that takes into account the entrance of Lithuania to the Nord Pool market and a Workday variable). β_2 controls for any change that affects daily prices and is changing over time in both experimental and non-experimental years. β_4 controls for anything that affects daily prices, that differs between markets, that differs between experimental/non-experimental years and that is constant within the years. β_5 controls for everything that changes over time in the experimental year. β_6 controls for any factors that affect daily prices across time for both markets but are constant across experimental and non-experimental years. β_7 controls for differences between Spanish and Nord Pool markets in the experimental year and are constant over time.

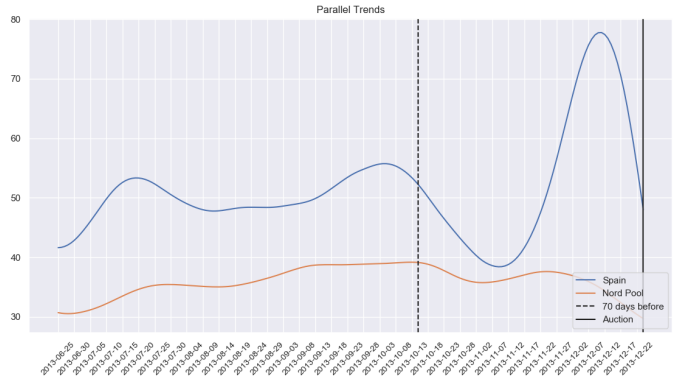
The third-level interaction β_8 captures the DDD estimation, i.e. the variation in prices specific to the treatment market in the experimental year, 70 days before

the auctions (relative to the control market, non-experimental year and untreated period). The key identifying assumption for this interpretation is that the change in prices is an unbiased estimate of the counterfactual. While we cannot directly test this assumption, we can test the hypothesis of parallel trends in the control and treatment markets during pre-intervention periods and during the non-experimental year. If the trends are the same in the pre-intervention periods, then we would expect the same pattern in the post-intervention period in both markets.

Further insights into this behavior are provided in Figure 2.5 where auctions number 24 and 25 are depicted. Note that we only have information from January 2013 onwards and so we cannot test the parallel trend assumption prior to the first auction. We use the Hodrick-Prescott filter to remove the cyclical component (Hodrick and Prescott, 1997). Treatment days lie between the dotted and solid lines. Before the dotted lines, we would expect to find parallel trends, and thereafter, divergent trends. As can be seen, in the pre-intervention period, prices in both markets present a similar pattern. However, immediately after treatment, while the pattern of prices in the Nord Pool market did not change, the prices in the Spanish market increased.



(a) Auction Number 24



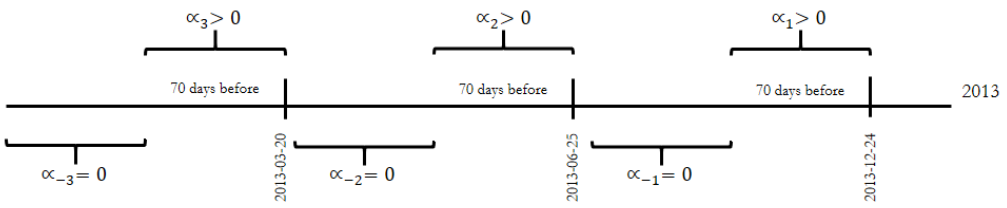
(b) Auction Number 25

Figure 2.5. Parallel Trend

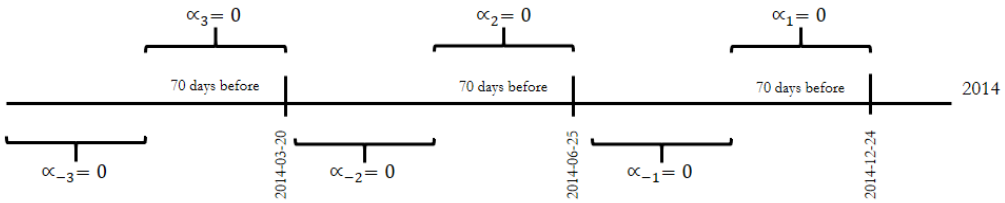
We formally test that the pre-intervention time trends and the non-experimental year for both the control and treatment groups are not different in the absence of regulation. This requires that the difference between the two markets is constant over time. In this model, we have multiple treated periods which makes it difficult to provide a simple visual inspection. One way to test the assumption of parallel trends is to evaluate the leads (LE) and lags (LA) of the treatment effect (Pischke, 2005). Thus, instead of a single treatment effect, we include m leads and q lags:

$$P_{ijt} = \beta_1 x_{ijt} + \beta_2 IP_t + \beta_3 \rho_j + \beta_4 T_i + \beta_5 \rho_j IP_t + \beta_6 IP_t T_i + \beta_7 \rho_j T_i + \sum_{t=-m}^{q} \alpha_t \gamma_j IP_t * T_i + \epsilon_{ijt} \quad (2.6)$$

where α_t is the coefficient of the tth lead or lag. In Figure 2.6, we illustrate what it is we are looking for. Thus, in 2013, we would expect the treatment effect to show a divergent trend (α_1, α_2, α_3 are positive). On the other hand, before the treatment effect we would expect to find a parallel trend (α_{-1}, α_{-2}, α_{-3} are zero). However, in 2014, we would expect to find a parallel trend across the whole period and, therefore, all γ coefficients should be zero.



(a) Year 2013



(b) Year 2014

Figure 2.6. Schematic Representation of Leads and Lags

Our results are reported in Table 2.10: The key assumption of parallel trends before the treatment effect is fulfilled. Again, we should stress that we only have information from January 2013 onwards and we cannot, therefore, test the parallel trend effect prior to the first auction.

Parallel Trends	coef
LA_trend_3_2013	-
LE_trend_3_2013	0.2161** (0.11)
LA_trend_2_2013	-0.0586 (0.084)
LE_trend_2_2013	0.1699*** (0.046)
LA_trend_1_2013	0.0921 (0.064)
LE_trend_1_2013	0.1257** (0.058)
LA_trend_3_2014	-0.047 (0.109)
LE_trend_3_2014	-0.1351 (0.083)
LA_trend_2_2014	-0.1837 (0.121)
LE_trend_2_2014	-0.0484 (0.051)
LA_trend_1_2014	0.0442 (0.043)
LE_trend_1_2014	0.1182 (0.114)

Table 2.10. Parallel Trend Estimation

In the methodology section, we explain why in this particular context difference-in-difference-in-differences estimations (DDD) are the most appropriate. There, we stressed that to employ this methodology two conditions must be met: Namely, (1) the differences in prices during 2013 must be related to the auctions, and (2) there must be no differences in prices in the counterfactual year.

Let's begin with the first condition. Using only data for 2013, we can evaluate a difference-in-differences estimation in which the treated group is the Spanish electricity market ($T_i = 1$) and the control group is the Nord Pool electricity market ($T_i = 0$). Recall, that in 2013 there were three auctions. Here, we seek to determine if the introduction of these auctions artificially raised the market price before they were held. If this is found to be the case, condition (1) is fulfilled.

From the results derived above, we use a 70-day time window before each auction. Therefore, our treatment variable is equal to 1 in those periods ($IP_t = 1$), and 0 otherwise ($IP_t = 0$). The difference-in-differences estimator is thus the coefficient of $IP_t * T_i$. As can be seen in Table 2.11, the results show a significant increase in prices before the CESUR auction in the Spanish market when compared to prices in the Nord Pool market. We can, therefore, conclude that condition (1) is fulfilled.

The gap between the Spanish market and the Nord Pool market increased on average 11%, 70 days before each auction in 2013.We can then conduct the same analysis but this time for 2014, the year in which the auction was eliminated as a regulatory mechanism. If condition (2) is fulfilled, we would expect no increase in prices during those same periods in which auctions were held. The results, reported in Table 2.12, confirm that there was no significant increase in prices before the CESUR auction in the Spanish market compared to prices in the Nord Pool market in 2014. Thus, condition (2) is also fulfilled.

Difference-in-Difference (2013)	P_{it}
$IP_t * T_i$	0.1047** (0.051)
FE	i, t
AR	Yes
Instrumental Variables	Yes
Control Variables	Yes
Trend	$i * period$

Table 2.11. Difference-in-Differences Estimation: Year 2013

Difference-in-Difference (2014)	P_{it}
$IP_t * T_i$	-0.0442 (0.032)
FE	i, t
AR	Yes
Instrumental Variables	Yes
Control Variables	Yes
Trend	i*period

Table 2.12. Difference-in-Differences Falsification Test

2.6.3 The Effect of Mandated Auctions on Prices

Table 2.13 summarizes the averages values of the triple differences estimation (DDD). The top panel compares the average prices recorded in the Spanish market in the experimental year (2013) with those in the non-experimental year (2014). Each cell contains average values, standard errors and number of observations. There was a 2.68-euro increase in price in the periods before the auction over the 2013 price, compared to an 8.88-euro price fall in 2014. There was an 11.56-euro relative increase in the Spanish market price before the auctions (the difference-in-difference effect). However, if there was a distinct market shock in Spain, this estimate does not identify the true impact of the auctions. In the bottom panel, we perform the same analysis for the Nord Pool market. We find an increase in the 2013 price relative to that for 2014 of 3.29 euros, considerably lower than that recorded in the Spanish market. If we consider the difference between the treatment and control groups, then there was an 8.47-euro increase in the relative markets in the year when auctions were held, compared to the change in relative price recorded in 2014.

However, the comparisons in Table 2.13 do not take into consideration other sources of variation in these price differences. As we explained earlier, to obtain the difference-in-differences between the Spanish and Nord Pool market, we must control for factors that vary over time and between markets. Differencing those factors will result in more accurate estimations. In Table 2.14 we present the equation (2) estimation results. Columns (1) to (6) contain data from the Spanish and the Nord Pool market in 2013-2014, and columns (7) and (8) summarize the double difference analysis in 2013 and 2014, respectively. Column (1) shows the regression when no autoregressive terms are considered. Serial correlation is very present in electricity prices (see Fezzi, 2007 for a detailed analysis) and we cannot therefore rely on a model that does not take these factors into account (see Appendix B: ARMAX Model-Building Process section). When we take the simplest model which also considers the autoregressive terms, our results drastically change. The estimated effect of the mandated auction on prices is statistically significant at a 5% level and it represents a relative increase of 12% in the treated market (the Spanish market compared to the Nord Pool market) during the experimental year (2013). Column (3) does not consider control variables (as we mentioned earlier, we are including the Oil Price, Gas Price, Unavailable Power and the entrance of Lithuania to the Nord Pool Market and weather condition variables). The effect on prices drop to 9% but is still significant at 5% level. Columns (4) and (5) do not incorporate an instrumental variable estimation for quantities and fixed effects (see Appendix A: Instrumental Variables Tests section), respectively. Both show similar results with a net increase on prices around 16.5-16.7%.

Market / Year	IP=0	IP=1	Policy Difference in market
A. Treated Market: Spain			
2013	41.97	44.65	2.68
	(17.69)	(17.54)	(24.91)
	[137]	[228]	
2014	47.40	38.52	-8.88
	(12.69)	(16.67)	(20.95)
	[136]	[228]	
Year difference in market	-5.43	6.13	
	(21.77)	(24.20)	
Difference-in-Difference	11.56		
	(32.55)		
A. Treated Market: Nord Pool			
2013	37.54	38.44	0.90
	(6.24)	(5.42)	(8.27)
	[137]	[228]	
2014	31.11	28.72	-2.39
	(3.41)	(4.56)	(5.70)
	[136]	[228]	
Year difference in market	6.44	9.72	
	(7.11)	(7.08)	
Difference-in-Difference	3.29		
	(10.04)		
DDD	**8.27**		
	(34.07)		

Table 2.13. Average DDD values. *Note: Cells contain mean log daily price for the group identified. Standard errors are given in parentheses; sample sizes are given in square brackets.*

	DDD Estimation						**DD 2013**	**DD 2014**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
$IP_t * T_i * \rho_j$	**0.02120**	**0.1134****	**0.087****	**0.1543*****	**0.1529*****	**0.1404****		
	(0.09)	(0.048)	(0.04)	(0.056)	(0.057)	(0.055)		
T_i	-0.00120	0.0871**	0.01930	-0.1629***	-0.1268***	0.0969***	-0.2044***	0.24930
	(0.064)	(0.037)	(0.024)	(0.062)	(0.037)	(0.035)	(0.076)	(0.6)
IP_t	0.010	0.0686***	0.0154*	0.059***	0.0466***	0.0434***	-0.0226	0.02460
	(0.024)	(0.023)	(0.009)	(0.021)	(0.015)	(0.015)	(0.012)	(0.02)
ρ_j	0.0020	0.2333***	0.1013***	0.1887***	0.1373***	0.1478***		
	(0.043)	(0.053)	(0.029)	(0.041)	(0.036)	(0.036)		
$IP_t * \rho_j$	-0.0080	-0.0523**	0.00780	-0.0792***	-0.0689***	-0.0654***		
	(0.027)	(0.022)	(0.013)	(0.023)	(0.019)	(0.018)		
$T_i * \rho_j$	-0.01890	-0.1173***	-0.0967**	-0.0586*	-0.072**	-0.0666*		
	(0.059)	(0.043)	(0.039)	(0.035)	(0.036)	(0.036)		
$T_i * IP_t$	0.01330	-0.0713*	-0.03510	-0.0673*	-0.05480	-0.05040	0.1047**	-0.04420
	(0.057)	(0.039)	(0.035)	(0.035)	(0.036)	(0.034)	(0.051)	(0.032)
AR	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CONTROL	Yes	No	No	Yes	Yes	Yes	Yes	Yes
IV	Yes	No	Yes	No	Yes	Yes	Yes	Yes
FE	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Obs	1452	1452	1452	1452	1452	1452	724	728
R2	0.069	0.579	0.621	0.648	0.616	0.641	0.648	0.642

Table 2.14. Triple Differences and Difference-in-Differences Results of The Effect of Mandated Auctions on Prices

Column (6) incorporates autoregressive terms, control variables, a quantity estimation using instrumental variables and fixed effects. The coefficient of $IP_t * T_i * \rho_j$ indicates that the relative price increased by around 15.07 percent in the treated market during the experimental year. These estimates are highly significant and present more robust standard errors compare to a difference-in-differences model (because of the double difference).

Columns (7) and (8) show the difference-in-differences results for 2013 and 2014, respectively. As we mentioned previously, two of our identification assumptions were that the coefficient on the $T_i * IP_t$ should be positive for 2013 and not significant for 2014. Column (7) reveals that the interaction is 11% and is significant at the 5% level. And column (8) reveals that this interaction is statistically non-significant. However, a double differences model only controls for unobservable factors that affect daily prices and are different between markets, but are constant over time. The triple differences models presented additionally control for unobservable factors that cause differences in daily prices for both markets to change over time.

Every column with autoregressive terms shows a similar outcome. The results confirm the conclusion that before the CESUR auctions the Spanish market experienced significant price increases. Specifically, we find a total increase in the difference of Spanish and Nord Pool markets between 9 and 17% when CESUR

auctions were introduced. If the objective of mandated auctions is to enhance competition, these results seem to contradict the initial assumption and what was stated empirically and theoretically by several authors (Allaz and Vila, 1993; Green, 1999; Wolak, 2000; Woo et al, 2004; Strbac and Wolak, 2017; Wolak, 2017). Because mandated auctions do not merely affect their own market, it requires careful evaluation when regulators are considering improving assignative and productive efficiencies.

We acknowledge that the model has two important limitations. First, we do not have data before 2013 for the Nord Pool market, which could enhance our results and make them more robust with a large period and a large number of auctions. Second, microdata about CESUR auctions is incomplete and unavailable, particularly related to the other mechanism in which firms may have affected prices by retiring quantity offers. This could enrich the analysis and explain some of the differences in CESUR prices.

2.7 Conclusions and Policy Implications

Natural concentration and price volatility are two common characteristics of liberalized electricity markets. As such, regulation should serve as a tool for delivering an essential service at an affordable price. In accordance with this line of thinking, several authors have argued that the introduction of fixed-price forward contract obligations increases competition and leads to prices that are closer to marginal costs. Auctions are usually used as a competitive mechanism to get both efficient supply and a competitive price. However, even though the auction that assigns the contracts is competitive enough, bidders who have market power over spot prices may influence the final outcome of the auctions by inflating price expectations. This chapter contributes to the literature on the results of the introduction of mandated auctions by studying the dynamics of differences in daily prices between two markets: one with regulation and the other without.

Between 2007 and 2013, the Spanish government held the *Contratos de Energía para el Suministro de Último Recurso* (CESUR) auctions. The auctions were long-term contracts for a fixed quantity at a price determined by a descendant price auction. This solution emerged as a way to foster liquidity in long-term markets and to stabilize the consumers' tariff cost. Comparing this market with the Nord Pool market (a control group that did not apply this regulation during the same period), we derive the following findings: (i) empirical evidence supports that daily prices suffered anomalous increases in days surrounding auctions, specifically, during the last 70 days before each auction. (ii) Using a combination of different methods we

find that electricity market prices rose by approximately 15 percent in the year when regulation was introduced in the Spanish market compared to the Nord Pool market. A number of factors lead us to conclude that the link between this regulation and the price increase is causal. First, the treatment and control groups presented similar trends in the pre-intervention period and second, these two groups presented similar trends once the auction was no longer operational.

Electric companies had several incentives to get higher CESUR prices. Besides risk volatility, they received payments and discounts on the energy supplied in this market. Clear evidence of ex-post forward premiums is presented in Peña and Rodriguez (2018), where they find a total premium average of 7.22%. However, the mechanisms by which firms could have got higher prices were not closely studied. Firms had two alternatives to influence prices (Fabra and Fabra Utray, 2012): (i) by taking off their supply offers during the auction (and therefore reducing competitive pressure), and (ii) by affecting parallel market expectations. As a consequence, we focus on the second mechanism as the first mechanism is hard to analyze due to incomplete and unavailable microdata about the CESUR auctions.

The theoretical model developed herein supports our main conclusions. We derive two propositions for the above result: First, the inherent characteristics of markets of this kind serve as an incentive to collusion. In other words, there is a threshold at which the set of collusive prices in both markets is a sustainable equilibrium of the perfect subgame. The natural reaction, therefore, is to avoid pro-competitive regulations. Second, expected exogenous price shocks, which originate once the auction is held, generate a perfect subgame equilibrium where prices are higher than without it. Conceptually, this suggests that a fixed tariff in a market characterized by high volatility – as is the case of the electricity market –induces firms to seek to minimize adverse outcomes by raising their prices.

Our main policy conclusion is that the uncompetitive outcome was originated in the deficient design of this particular regulatory mechanism. Well-designed mechanisms need to take into account the specific characteristics of the electricity market. As argued at the outset to this chapter, the power of Spanish firms to react seems to stem from anti-competitive structures, a low level of interconnectivity – with just two firms concentrating 64% of generation capacity, and a pivotal index rate below 110% for more than 5% of the time when the regulation was introduced. Moreover, liberalization exposed the electricity market to high price volatility. A format of repeated auctions and fixed prices in an environment of natural concentration and high volatility in prices seem to be the main reasons why firms colluded (either to protect the inherent characteristics of the market or to absorb the volatility risk). Additionally, some particularities of the auction design may have discouraged competition. Especially that they knew the ranges of supply excess from where they

could have intuited their pivotal power. Consequently, regulators who are interested in limiting market power should mainly be focused on the macro and micro aspects of this kind of regulation.

We can think of two different alternatives to improve competitive results. First and most obvious is that market regulators may consider deregulating the long-term contract market. However, removing mandated auctions and liberalizing the sector when the market has severe concentration problems and low levels of interconnectivity does not seem reasonable. Electricity demand is inelastic per se, therefore, liberalizing where there is no competition may yield to worsening uncompetitive equilibriums. This is explained by Green (2004), who shows that retail competition when price volatility is high disincentives long-term contracts and produces higher prices. Furthermore, international empirical evidence is still inconclusive regarding this point. Mobility between companies tends to be very low and static due to contract complexities and high switching costs. Additionally, prices between incumbent companies and competitors do not converge in many cases, and price discrimination by region is a common practice. Moreover, companies that usually take control are incumbents from other industrial sectors or from other regions (see Defeuilley, 2009 for a detailed analysis on retail competition evidence).

Second, instead of deregulating in order to obtain a competitive situation, where consumers can choose between different contract options, an alternative suggested by some authors is mandatory basic electricity services -wholesale price plus a regulated margin- (see Joskow, 2000). In theory this would provide a competitive benchmark where consumers could compare options. This mechanism is simple and transparent, and works as an alternative for the consumer when other options have higher prices than the market price. However, more research is needed to support this statement. Mandatory basic electricity services are based on the assumption that electricity markets are competitive enough to be a representative benchmark. But, as we stated in our main results, it would be hard to expect that retailers (which in the Spanish market are in general holding groups or subsidiaries from companies that are present in the generation market) do not seek mechanisms to inflate expectations. An intermediate solution could be the introduction of "Price to Beat" policies (variable prices adapted to the generation cost). In this case, incumbent companies are only allowed to offer a lower price than the price to beat if, and only if, a certain amount of time has passed or after a share of residential and small business customers are served by other suppliers (Adib and Zarnikau, 2006 and NERA, 2007). Whatever the alternative, what is clear is that getting completely rid of regulation frameworks leaves market operators with very limited control over concentrated markets. Liberalizing where no competition exists is not a prudent decision. We want to emphasize that our study does not seek to be critical of the mandated auc-

tion regulation, quite the opposite: to focus attention on the problem that lies in the deficient design of this regulatory mechanism and not in the regulation, per se.

Appendix A: Instrumental Variables Tests

As discussed above, a problem that can arise with the OLS estimation is the presence of endogeneity in the quantities of energy (i.e., q_t is correlated with μ_t). This means we cannot infer whether the changes in price and quantity are due to shifts in demand or supply. Wright (1928) suggested that this problem can be addressed by using curve shifters, i.e. if we find instruments that are related to the demand conditions but which do not affect the cost function, then we would be able to identify the supply correctly.

To do so, we need instruments Z that satisfy two conditions: First, they need to be correlated with the endogenous variable ($cov(Z, q) = 0$) and, second, the instruments must be orthogonal to the error term ($cov(Z, \mu) = 0$). The second condition cannot be empirically proven (as we do not know μ); thus, the validity has to be left to economic reasoning. To test the first condition, we can express the reduced form equation of q_t by conducting a two-stage least squares estimation. Therefore, in the first stage we can construct an equation such as:

$$q_t = \pi Z + v_i$$

And using the predicted values $\hat{q}_t$, we can run a second stage as:

$$y_t = \beta X_t + \gamma IP_t + \alpha \hat{q}_t + \mu_t$$

As $\hat{q}_t$ is uncorrelated with μ_t, the classic endogeneity assumption remains.

Specifically, we propose using a working day dummy variable as our instrument (which takes a value of 0 on weekends and national holidays, and 1 otherwise). Thus, the assumption we make is that working days are correlated with the quantities demanded but do not directly affect the prices offered.

When we use a vector of instruments z we expect them to satisfy: (1) z is uncorrelated with the error term; (2) z is correlated with the instrumental variable; (3) z is strongly correlated with the instrumental variable. The first condition invalidates the instrument, the second condition makes the instrument irrelevant, and the third condition makes the model weakly identified.

We use a two-stage least squares estimation to evaluate these conditions. The first requirement is difficult to evaluate because we cannot observe the error term. But we can test the second requirement using a reduced form with all the exogenous

variables. As can be seen in Table 2.15, both the quantities for Spain and those for Nord Pool satisfy condition (2). Similarly, the F-statistics of the instrumental variables (under the null hypothesis that all instrument coefficients are null) validate the third condition. However, serial correlation, non-stationarity and non-normality problems may persist.

First Stage	coef	std error	t	$P > \|t\|$	[0.025	0.975]	F-statistic	F-value
Workday Spain	0.1469***	0.013	11.483	0.00	0.122	0.172	85.49	0.00
Workday Nord Pool	0.0636***	0.011	5.918	0.00	0.042	0.085	18.92	0.00

Table 2.15. Reduced Form $q = \gamma z + \beta_1 x_1 + \beta_2 x_2 \ldots + \mu$

Appendix B: ARMAX Model-Building Process

As we validated the instruments used, the classic endogeneity assumption remains. However, serial correlation, non-stationarity and non-normality problems could persist.

The autoregressive moving average (ARMA) is a time-series model that uses two polynomial terms to describe stationary stochastic processes: autoregressors (AR) and moving averages (MA) (Box et al., 2015). If, in addition, we include exogenous input terms, we obtain the ARMAX model.

The primary advantage of ARMAX models is that they allow us to correct the effects of serial correlation (especially present in electricity market prices) which may invalidate the estimations. However, models of this kind rely on very strong assumptions that have to be proved before any conclusions can be drawn. They include:

1. The dependent variable series must be stationary.
2. The residual series must not exhibit serial correlation.
3. The estimated coefficients of exogenous variables must be significantly different from zero.
4. The sign of coefficients must be reasonable.
5. Exogenous variables must not exhibit multicollinearity.

Figure 2.7 shows the ARMAX model building approach proposed by Andrews et al. (2013) and based on the practical approach developed originally by Box and Jenkins (1970). The extensions applied to ARMAX models involve the introduction

of new exogenous variables that may disrupt the white-noise pattern of the residuals and which, therefore, might change the order of the autoregressive and moving average terms.

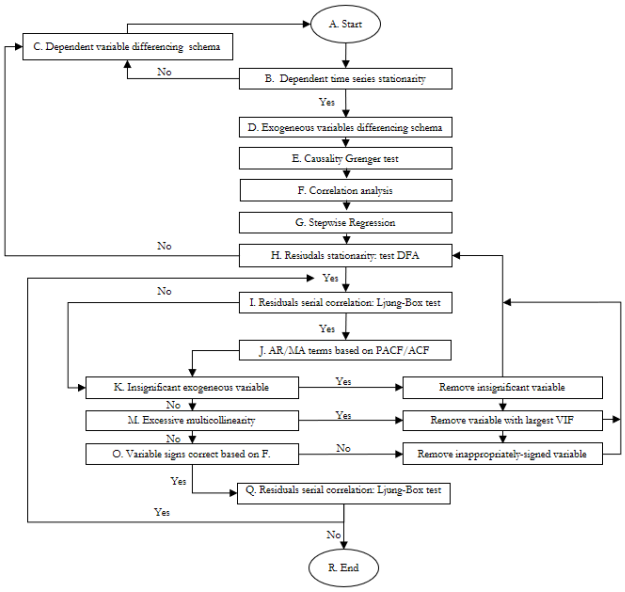


Figure 2.7. Armax Building-Model Process

The first assumption of the ARMAX model is that the dependent variable is stationary, i.e., the mean and variance do not change over time. We use an Augmented Dickey-Fuller (ADF) test (Fuller, 1976; Dickey and Fuller 1979) to evaluate this assumption. Under the null hypothesis, the time-series is a unit root, i.e., it is non-stationary.

Table 2.16 reports the ADF test result. We apply second differences to make the dependent variable time-series stationary. This differencing schema is also applied to exogenous-variable candidates, which permits correlations to be more stable over time.

$log(P_{Spain}/P_{NPool})$	**Diff(0)**	**Diff(2)**
Test Statistic	-2.99	-12.29
p-value	0.04	0.00
Lags Used	14	12
Number of Observations Used	381	381
Critical Value (1%)	-3.45	-3.45
Critical Value (5%)	-2.87	-2.87
Critical Value (10%)	-2.57	-2.57

Table 2.16. Augmented Dickey-Fuller Test

We use a forward stepwise regression procedure to identify the most relevant

exogenous-variable candidates (see Table 2.7). This method seeks to fit a subset of attributes by successively incorporating new variables while evaluating the performance. Those attributes that improve performance can then be maintained definitively.

The ARMAX model works under the assumption that the residuals are white noise, i.e., a random sequence that cannot be predicted or, what amounts to the same thing, that the residuals are stationary and do not exhibit significant serial correlation. Serial correlation implies that the error term observations are correlated. Patterns in the error term could bias the significance of the exogenous variables. As shown in Table 2.17, the ADF test provides evidence that the residuals are stationary.

ADF Test - Residuals	
Test Statistic	-11.00
p-value	0.00
Lags Used	6
Number of Observations Used	314
Critical Value (1%)	-3.45
Critical Value (5%)	-2.87
Critical Value (10%)	-2.57

Table 2.17. Residuals Stationarity Test

To test serial correlation, we use the Ljung-Box test (Ljung and Box, 1978) which checks under the null hypothesis if the data are independently distributed[8]. The p-values from the Ljung-Box test (see Table 2.18) support the rejection of the null hypothesis. This is an indicator that we must add appropriate combinations of AR/MA terms which are identified from partial autocorrelation and autocorrelation functions, respectively.

	L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**L9**	**L10**
QLJB	77.45	80.33	90.65	94.25	94.33	94.62	97.59	106.64	109.28	109.76
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.18. Residuals Serial Correlation Test

The autocorrelation function (ACF) describes the correlation between observations and observations at a prior time step. The partial autocorrelation function (PACF) is the autocorrelation between observations in a time series with observa-

[8] As it tests the overall randomness based on the number of lags, we use the rule proposed by Hyndman and Athanasopoulos (2018) where lags are equal to $min(10, T/5)$ for non-seasonal time series.

tions at prior time steps but with the relationships of intervening observations removed. If autocorrelation exists, then there is some information that could explain the movements of the dependent variable but we are not capturing it.

We use the Bayesian Information Criterion (BIC) (Schwarz, 1978) to select the optimal combination of AR/MA terms. The BIC is based on a likelihood function which also introduces a penalty term to avoid overfitting[9].

Based on this criterion, we add seven AR terms and no MA terms. Our model is therefore an ARMAX(7, 0). Figure 2.8 illustrates the ACF and the PACF after the introduction of the autoregressive terms. This shows that autocorrelation and partial autocorrelation have been removed.

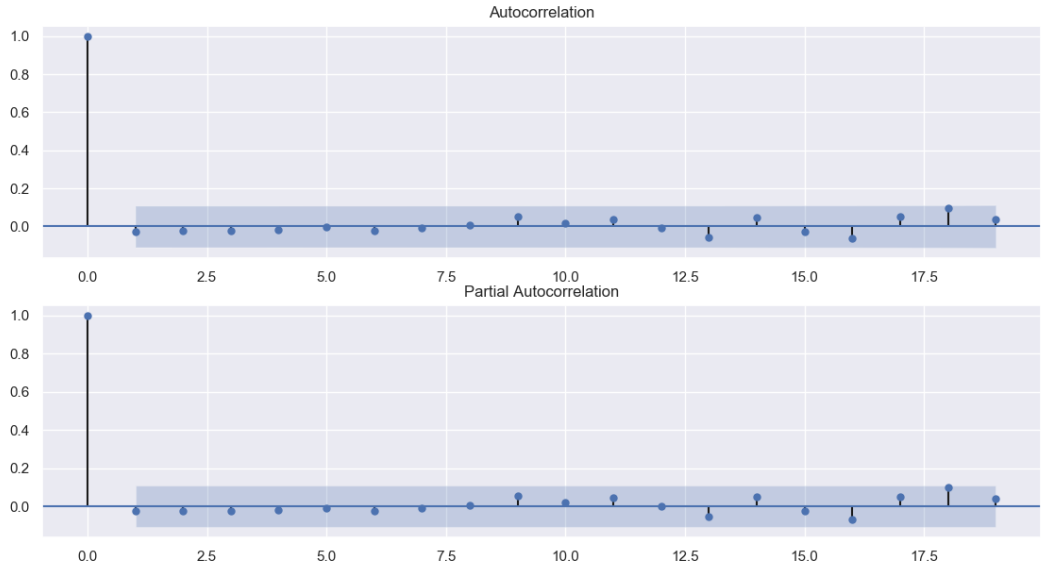


Figure 2.8. Partial Autocorrelation and Autocorrelation Functions

Additionally, in Table 2.19, we present the Ljung-Box test with the new configuration which also reveals strong evidence of no serial correlation.

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
QLJB	0.23	0.38	0.55	0.62	0.62	0.81	0.81	0.88	1.82	1.93
p-value	0.63	0.83	0.91	0.96	0.99	0.99	0.99	0.99	0.99	0.99

Table 2.19. Residuals Serial Correlation Test with ARMAX(7, 0)

The next assumption in the ARMAX building procedure is that the estimated coefficients must be significant. Verification of this is provided by the t-statistics

[9] $BIC = ln(n)k - 2ln(\hat{L})$ where $\hat{L}$ is the maximized value of the likelihood function of the model M, n are the number of data points and k the number of variables.

in the regression output. As shown in Table 2.20, every variable is still significant at a 95% confidence level. The assumption that the sign of the coefficient for each variable must be reasonable also seems to be satisfied.

	coef	**std err**	**z**	$P > \|z\|$	**[0.025**	**0.975]**
Nord Pool Production	3.8710	0.445	8.706	0.000	3.000	4.742
Spain Production	-2.3440	0.271	-8.637	0.000	-2.876	-1.812
Spain Imports	0.2538	0.054	4.697	0.000	0.148	0.360
Dummy Null Price	-1.6598	0.484	-3.426	0.001	-2.609	-0.710
Spain Production -1	0.5664	0.238	2.382	0.018	0.100	1.032
Portugal Production	-1.0529	0.270	-3.906	0.000	-1.581	-0.525
ar.L1	-0.9185	0.059	-15.602	0.000	-1.034	-0.803
ar.L2	-0.7691	0.077	-10.046	0.000	-0.919	-0.619
ar.L3	-0.5345	0.089	-6.028	0.000	-0.708	-0.361
ar.L4	-0.4354	0.089	-4.890	0.000	-0.610	-0.261
ar.L5	-0.3639	0.089	-4.068	0.000	-0.539	-0.189
ar.L6	-0.2941	0.077	-3.824	0.000	-0.445	-0.143
ar.L7	-0.1983	0.061	-3.233	0.001	-0.319	-0.078

Dep. Variable:	$log(P_{Spain}/P_{NPool})$	**Obs:**	321
Model:	ARMAX(7, 0)	R^2	0.6422
Durbin-Watson	2.02	**BIC**	456.192

Table 2.20. ARMAX Model Results

Next, we test the assumption that the surviving exogenous variables do not exhibit a significant level of multicollinearity, since this could cause standard errors to become too large (i.e., to overstate the p-values). We use the variance inflation factor (VIF) [10] to check if the variables are linearly independent. This measures just how much the variance of an estimated regression coefficient is increased as a result of collinearity. A rule of thumb threshold is that the VIF must be lower than 10, which is equivalent to each independent variable's variation being less than 90 percent explainable. Table 2.21 shows that multicollinearity is not a problem in the chosen exogenous variables.

Appendix C: LSTM Specification

Specifically, the long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) is a type of recurrent neural network that is useful for identifying long-term dependencies. Instead of using neurons, LSTM uses memory blocks connected

[10] $VIF = 1/(1 - R^2)$

Features	VIF
Nord Pool Production	2.70
Spain Production	2.33
Spain Imports	1.78
Dummy Null Price	1.04
Spain Production - 1	1.30
Portugal Production	2.44

Table 2.21. Variance Inflation Factor Index

through layers. A block has different gates that are used to manage the states and outputs. The input vector and the output from the previous step pass through a memory cell (using a sigmoid activation function) that retains any relevant information from the new input and forgets irrelevant information (if any) from the past. The forget gate resets the memory contents when they become irrelevant and the output gate applies a sigmoid activation function to the memory cell output (see Figure 2.9).

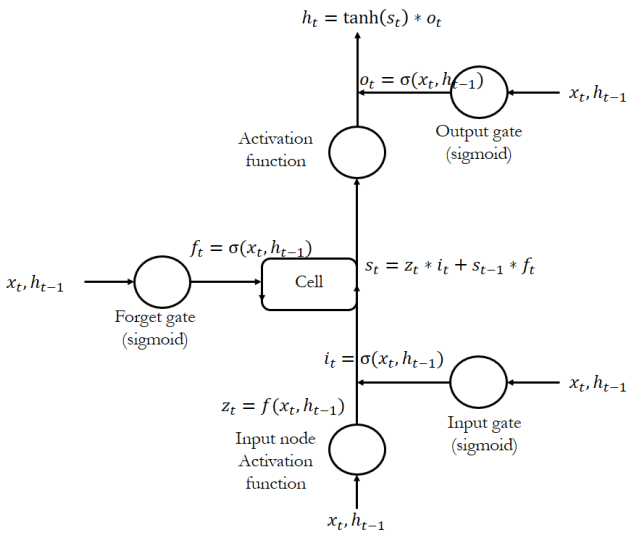


Figure 2.9. Schematic Diagram of the LSTM unit with forget gates

In this case, we use LSTM as an anomaly detection model, training only with normal data (here the year 2014 provides us with our training data) and using the prediction error as an outlier indicator. The prediction error of the normal data is constructed as a Gaussian distribution from which we derive mean and variance using a maximum likelihood estimation. On the test data, which contains normal and abnormal data (year 2013), we calculate the log probability densities of errors. Lower values indicate a greater likelihood of the observation being an anomaly.

By definition, we are using a stacked LSTM (i.e., comprising several layers) with

memory between batches. This allows us to identify when the state of the network is reset by using the final state of the sample i of the current batch as the initial state for i the sample of the next batch. Maintaining the state is crucial in data that present a long repeated pattern, because if not, the model will be influenced mostly by recent observations. Therefore, we maintain the state over all the training batches, and reset it only when the next batch is about to start, tuning the weights over different epochs, and allowing the LSTM to store relevant activations in the cell state.

As discussed, for training purposes we use the year 2014, which we assume to be a normal period. In this way, our test data correspond to 2013, for which we have anomaly data (70 days before each auction) and normal data (remaining observations). Based on this definition, we proceed to employ a typical semi-supervised outlier model. Our recurrent neural network comprises three recurrent layers each made up of 100 LSTM units, with a final dense output layer with one neuron that uses a linear activation function. In order to prevent overfitting, we use two different regularization methods: a dropout of 20% and early stopping set to 2 epochs. The dropout is a regularization method that randomly drops out a number (or percentage) of layer outputs. It experiments with different architectures of the same network in parallel, which becomes a more robust model as the training process is noisy and nodes are forced to actually learn a sparse representation (Srivastava et al., 2014). Additionally, we use early stopping which provides a rule of how many iterations we can run before the learner overfit (i.e., when the test loss starts to be worse than the training loss).

We use an Adagrad optimizer (Duchi et al., 2011) to adapt the learning rate to the parameters. This is achieved by performing smaller updates for parameters associated with frequent features, and larger updates for parameters associated with infrequent features. The main advantage of this optimizer is that it works well with sparse data. As Dean et al. (2012) show, the method greatly improves the robustness of the stochastic gradient descent optimizer.

We set the loss function as the mean squared error, which is well suited to problems in which real-value quantities are predicted. Thus, we calculate the average of the squared differences between the predicted and the real values.

Activation functions define the output of neurons given the previous input. Each layer is trained by using the hyperbolic tangent activation function (unless the input and output gates use sigmoid functions), which has the advantage of being less likely to become stuck (as it strongly maps negative inputs to negatives outputs).

In our experiment, we use 50 epochs (i.e., the number of times we go through the training set) and a batch size of 50 steps per epoch (i.e., the sample used in each iteration).

As we proposed a 70-day time window before auctions as our anomaly pattern, we expect the model to fail to reconstruct the prediction error (fail, that is, in terms of reconstructing a multivariate normal probability density function similar to that derived from the training error vector). By using a valid-set, we are able to define an optimal threshold that separates anomalies from normal behavior. Finally, this threshold is applied to the test error vector where we expect lower values to indicate that the observation is more likely to be an anomaly.

3 Machine Learning Forecasts of Public Transport Demand: The Case Study of Buenos Aires, Argentina

3.1 Introduction

The use of smart cards is becoming increasingly popular on public transport services. They are especially convenient for users as they reduce their transactional costs and boarding times, whereas for bus companies, they enable them to plan their schedules more effectively, improve commercial bus speeds, while indirectly allow them to reduce personnel and maintenance costs. But the cards have a further advantage that has yet to be exploited: they provide massive amounts of information ranging from tariffs to GPS-generated mobility patterns. Such a rich seam of data, if mined properly, should have great policy implications for public transportation authorities, sector regulators, transport operators and other interested parties, as well as the public in general.

Many recent studies stress the potential of smart card data as a tool for transport management and planning (Blythe, 2004; Bagchi and White, 2005; Agard et al., 2006; Utsunomiya et al., 2006; Morency et al., 2007; Pelletier et al., 2011; Munizaga and Palma, 2012; Ma et al., 2013; Kusakabe and Asakura, 2014; Briand et al., 2017; Ma et al., 2017, Li et al., 2018). Here, a particularly interesting dimension of analysis is estimating demand given that smart cards record the commuting characteristics of each passenger, including data about travel dates, time of day, origin and destination, journey times, etc. Indeed, if transport authorities and regulators could exploit this demand information, they would be able to optimize the transport network as a whole. Several examples can be found: Dou et al. (2015) use smart card data from the public transport service in Queensland city to predict individual mobility patterns. Zhou et al. (2017a) analyze the responses of passengers to weather conditions by combining meteorological data with large-scale smart card

data in the city of Shenzhen. Zhou et al. (2017b) use information of card swiping times to estimate bus arrival more precisely in Beijing. Kumar et al. (2018) use several features collected from the smart card data in the Minneapolis transport integrated network to develop an innovative method for trip chaining. Huang et al. (2018) combine mobile phone data, subway smart card data and taxi GPS data from Shenzhen to predict real time urban travel demand. Zhao et al. (2018) utilize transit smart card records from the rail-based system in London to predict the next trip within a day and its attributes. Ingvardson et al. (2018) analyze passenger arrivals and waiting times at rail stations in Copenhagen using smart card data from the public transport system in Denmark. Zhang et al. (2018) fusion self-reported revealed preferences data with smart card data collected from the Guangzhou city metro system to forecast metro passengers' path choices.

To date, however, few studies have resorted to the use of smart cards for their input data to predict demand and elasticities. Those that have include Seaborn et al. (2009), Munizaga and Palma (2012), Tao et al. (2014), Tao et al. (2016). One of the most innovative examples is provided by Arana et al. (2014) who analyze smart card data to predict demand. They use multiple linear regression analysis to assess the impact of weather conditions on the number of trips taken and the underlying motives for these trips. However, it was hard to find literature related to fare impact on public transport demand using smart card data. Examples are de Grange et al. (2013) who estimate demand elasticities for the integrated system in Santiago de Chile and Wang et al. (2015) who analyze fare changes in the Beijing Metro to calculate price demand elasticity. Nevertheless, demand elasticity is a problem that economic literature has widely aboard using other sources (Just to mention some examples: Goodwin, 1992; Preston, 1998; Nijkamp and Pepping, 1998, Hensher, 1998; FitzRoy and Smith, 1998; Bonnel and Chausse, 2000; Hanly et al., 2002; Matas, 2004; Bresson et al., 2003, 2004; Balcombe et al., 2004; Paulley et al., 2006; García-Ferrer et al., 2006; Holmgren, 2007; Crotte, 2008; Graham et al., 2009; Albalate and Bel, 2010; Tsai et al., 2014, Tsai and Mulley, 2014).

Regarding time-series forecasts of public transport, quite number of authors apply Autoregressive Moving Average regressions (ARMA). This kind of model and its variations maintain the interpretation of traditional linear models but also take into account the residuals pattern. Examples are Ahmed and Cook (1979) who apply ARMA models to forecast traffic state in USA, Williams et al. (1998) who use SARIMA models in traffic seasonal flows, Suwardo et al. (2010) apply ARIMA models to predict bus travel time in Malaysia, Gong et al. (2014) who use a SARIMA model to predict the arrival passengers and the number of waiting passengers, Xue et al. (2015) who forecast short-term passenger demand based on smart card data from Shenzhen, China and Milenkovic et al. (2018) who use SARIMA

models to forecast railway passenger flows in Serbia. However, most analyses in the literature employ linear models, which means they make demand predictions using techniques that are better suited for finding causal relationships. Kremers et al. (2002) show that from 79 studies related to public transport demand, 55 models were estimated using log-linear models and 24 used linear models.

In this study, our focus is very clearly on methods that enhance prediction capabilities as opposed to the assessment of marginal effects on target variables. Moreover, traditional models fail to take into account the importance of performance in terms of out-of-sample errors (Mullainathan and Spiess, 2017). Their focus on in-sample data is not an optimal approach to forecasting demand. Here, in contrast, we highlight the importance of testing a prediction model with out-of-sample data from smart cards, given that the results of a forecast and the fulfillment of actual events may differ.

The objectives of this chapter can therefore be clearly stated. First, we present various supervised machine learning (SML) techniques for predicting public transport demand using smart card data. SML techniques train models using historical data in such a way that they learn from the patterns that emerge. However, more importantly, machine learning techniques are validated with test data so that their real predictive power can be determined. Second, we then compare these results to linear models outcomes to determine the gain in performance achieved with SML techniques. The comparative analysis is focused in three aspects: interpretability, predictive power and demand elasticity.

It follows, therefore, that machine learning (ML) algorithms are specifically designed for making predictions, while linear models are not (Xie et al., 2003; Zhang and Xie, 2008; Kleinberg et al., 2015; Zhao et al., 2018). Moreover, ML are able to exploit several data types and complexities. But perhaps their main advantage is the fact that computers can be programmed to learn from data, revealing previously hidden findings as they discover historical relationships and trends. ML techniques can improve the accuracy of predictions by removing noise and by taking into account many types of estimation, although not necessarily without bias. Moreover, ML allows for a wide range of data, even when we have more predictors than observations, and it admits almost every type of functional form when using decision trees, ensuring a large interaction depth between variables. Of course, the downside of ML techniques is biased coefficients; however, if our main concern is the accuracy of the prediction, then any concerns regarding biased estimators become irrelevant.

Thus, number of recent studies have suggested that machine learning algorithms are well suited to predict travel behavior (Omrani et al., 2013; Omrani, 2015; Hagenauer and Helbich, 2017; Golshani et al., 2018; Wang and Ross, 2018; Zhao

et al., 2018; Gu et al., 2018). There are several applications of machine learning techniques related to predicting traffic flow, traffic speed, travel time and travel behavior. For instance, Yu et al. (2011) compared several supervised models to predict bus arrival times and bus stops using survey data from Hong Kong. Pitombo et al. (2017) find that tree based decision models outperform traditional gravity models in predicting destination choices. Ke et al. (2017) analyze on-demand ride service data in Hangzhou city and try to forecast short-term passenger demand via a long short-term memory network. Liu and Chen (2017) also use supervised deep learning models but to predict the hourly passenger flow in Xiamen city. Similarly, Wu et al. (2018) utilize deep learning models to explore traffic flow prediction by exploiting detectors data from an interstate highway. Xu et al. (2018) forecast bike sharing travel demand using a deep-learning approach and data collected form Nanjing city. Sekula et al. (2018) compare diverse supervised models and predict hourly traffic volumes by combining data from automatic traffic recording and several data sources from Maryland city. Wang et al. (2019) predict traffic speed with a variety of supervised models, exploiting data collected from automatic vehicle identification detectors in Xuancheng city. Ma et al. (2019) use a combination of unsupervised and supervised algorithms to predict bus travel time by using vehicle trajectories and bus smart card data. Liu et al. (2019) compare several deep learning architectures to predict short-term metro passenger flow using data collected from the Nanjing city metro system. Other applications of machine learning techniques are related to driving safety and efficiency: Yuang and Cheu (2003) use several SVM classifiers to detect arterial and freeway incidents in peak periods in survey data from Singapore. Guo et al. (2018) fusion multiple supervised models to predict short-term traffic and capture traffic uncertainties such as incidents and accidents. Bejani and Ghatee (2018) propose a driving style evaluation system by comparing a variety of supervised algorithms based on smartphone sensor data. Similarly, Yi et al. (2019) use smartphone data to develop a personalized driving state recognition system using several supervised classifiers. In this way, given that the literature informs us that SML methods are better than traditional econometric prediction techniques, our hypothesis is that this should also be illustrated in transport demand predictions based on data retrieved from smart cards.

Here, we study the case of the Autonomous City of Buenos Aires (CABA in its Spanish acronym), the capital city of Argentina. We analyze the use of the SUBE (*Sistema Único de Boleto Electrónico*), a smart card employed on the city's public transport services (train, metro and bus) since 2009, and which has now been extended for use throughout the country. The card (similar in many respects to a credit card) collects multiple details about travelers and the journeys they make, and provides information about the location of vehicles using GPS and is, therefore, an

extremely rich and reliable source of information. This smart card data gives us a great opportunity to evaluate not only interpretability and predictive power but also demand elasticity. During the period analyzed, fares suffered three different nominal increases which permit us to evaluate short-term price elasticity.

This chapter makes two main contributions to the literature. First, we apply ML techniques to public transport data using smart card data in the particular case of CABA. These algorithms, of frequent application in other areas, when used in the transport sector, show substantial improvements on previous forecasts. Second, we compare interpretability power, we show the potential of ML techniques for making predictions in comparison to those obtained using traditional econometric estimations and we examine short-term demand elasticity with an increase in nominal fares.

In short, we present a broad overview of the comparative between SML automated tools and traditional time-series models for exploiting public transport smart card data, and, in so doing, we contribute to the discussion on the trade-off between accuracy and causality, and why this is fundamental for empirical predictions.

The remainder of this chapter is structured as follows. Section 2 presents the study case and the data collection. Section 3 summarizes the methodology applied. In Section 4, the results are shown regarding: interpretability, predictive power and demand elasticity. Finally, conclusions are provided in Section 5.

3.2 Case Study and Data

Public transport in CABA is provided in an integrated system that combines urban buses – starting and terminating within the city's limits – with suburban buses – starting (terminating) in CABA and terminating (starting) in another district, an incipient underground metro network and inter-city trains.

Figure 3.1 shows that 80% of all trips are made by bus, with the train operating primarily as a feeder from the suburbs. Given, therefore, the predominance of the bus as the main public transport service, the analysis we undertake herein focuses exclusively on this particular transport mode.

Figure 3.2 shows the routes taken by buses in real time one day in March 2016 at 08:00 a.m. On average, there are 1,271 travelers per hour per bus route and, within the analyzed time period, there are approximately 32 active bus routes. Each month, an average of 20-30 million trips are registered.

In general, CABA has a temperate climate, without any extreme temperatures. The city is sited on the Río de La Plata, and so humidity levels are quite high (around 70-80%). In summer, the average temperature oscillates around 25 C^{o} and

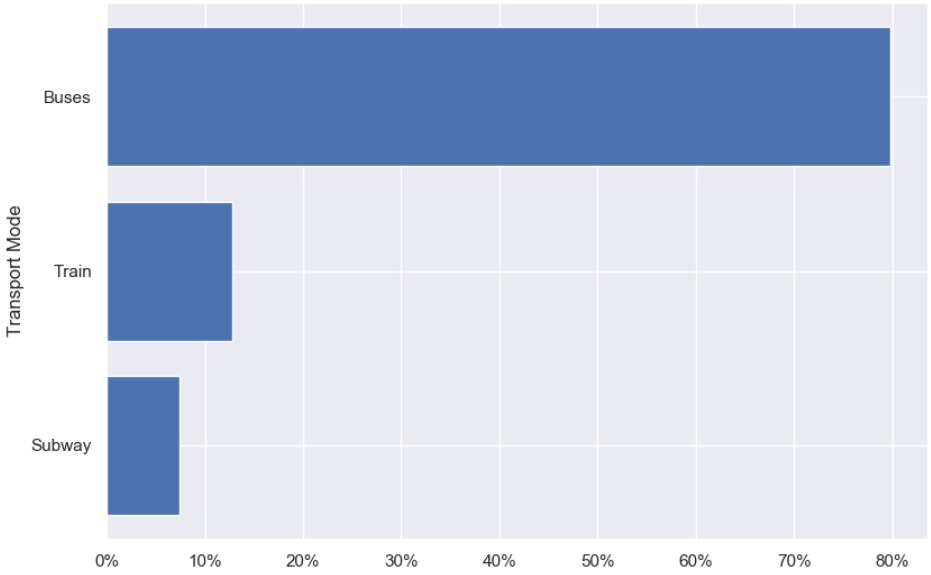


Figure 3.1. Domiciliary Mobility Survey (2013)

Figure 3.2. Location of CABA

in winter around 10 C°. Figure 3.3 shows that during midsummer rain is abundant.

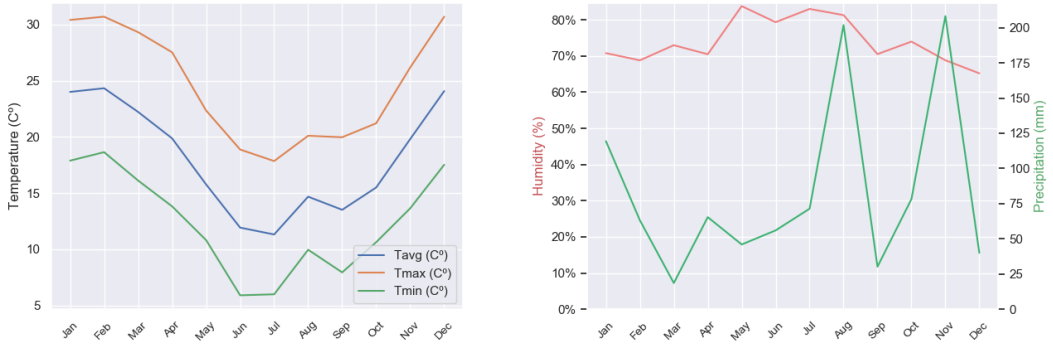


Figure 3.3. Weather Conditions in CABA (source INTA)

During the period from November 2013 to June 2016, the public system suffered three fare increases. The first increase was in January 2014, the second one in July 2014 and the last one in April 2016. As can be seen in Figure 3.4, the number of passengers follows a cyclical trend with low demand peaks during summer holiday periods (January-February). From this illustration, it is hard to conclude if there was a price effect on demand.

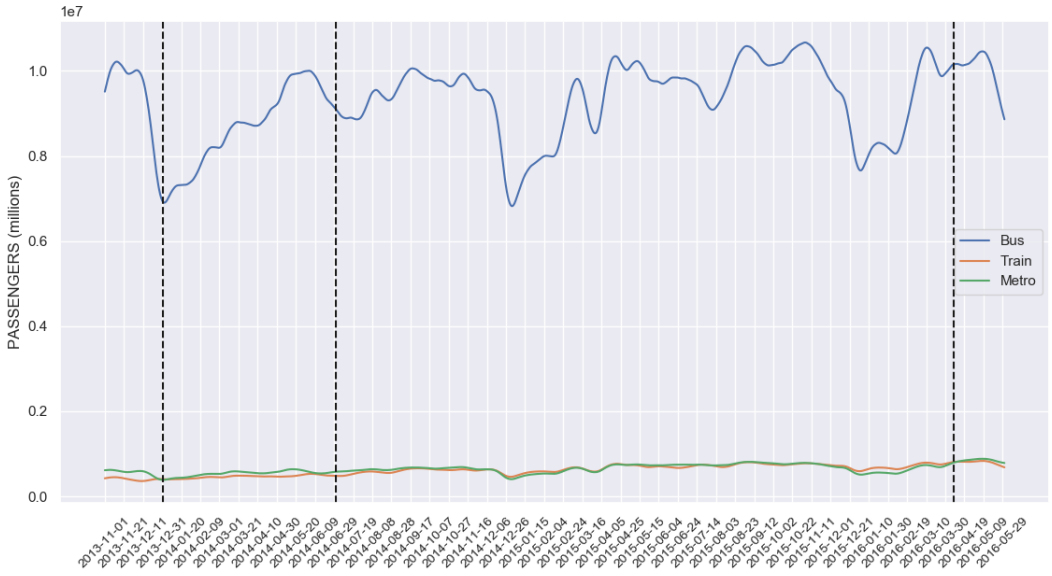


Figure 3.4. Number of Passengers by Day

In addition, as we are evaluating demand elasticity, it is important to remark that during the period analyzed there was a persistent inflation. If we deflate fares, there is almost no increase in prices. In fact, fare increases intended to recover the original

real value. For instance, the weighted average nominal increase in May 2016 was around 80%. However, the real increase was only 1-dollar cent. See Figure 3.5.

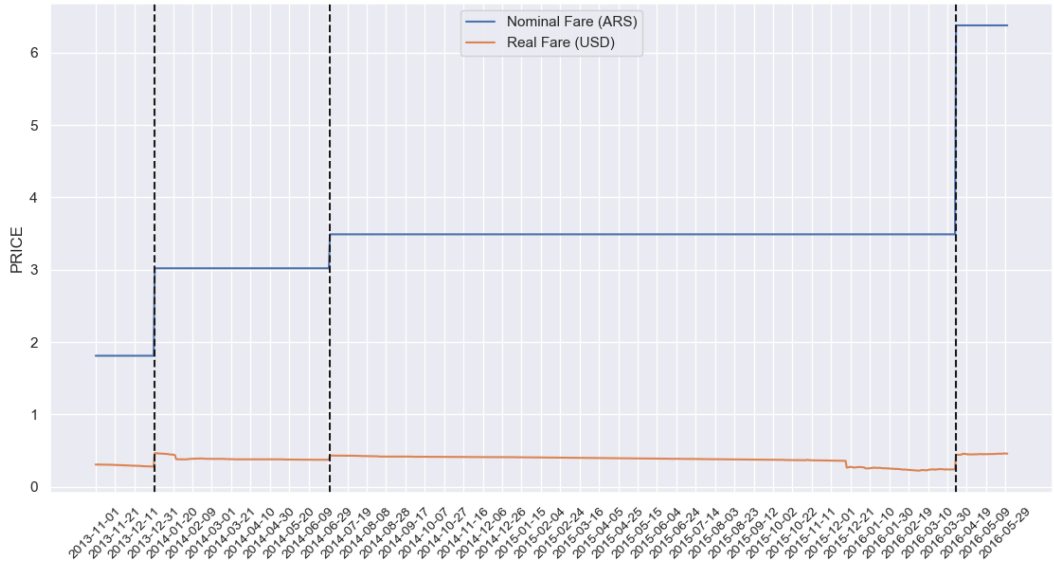


Figure 3.5. Nominal and Real Fares Evolution

This money illusion effect give us a great opportunity to evaluate the nominal demand elasticity. As the real effect was negligible, we would expect that demand have not decreased with higher nominal fares. Any other result should be considered part of a money illusion effect. The last part of this study is therefore going to be dedicated to analyze the demand behavior.

3.2.1 SUBE

We use data from the SUBE, sorted primarily by bus route and by hour during the period that extends from September 2013 to June 2016. It should be borne in mind that the bus system has undergone no modifications or developments since 2012, in accordance with Ministerial Resolution 422/12. This means that during the period analyzed, the system's supply did not vary in terms of routes, vehicles, frequency of services or quality (all buses being identical as they have to respect the technical specifications stipulated under provincial law). In addition, use of the SUBE is obligatory for all companies and it is not possible to use an alternative form of payment. These facts are significant with regards to potential problems of endogeneity, given that we seek to predict demand. Furthermore, note that competition between firms operating the system is meaningless because the government controls not only their costs but their profits too (as well as defining service quality and quantity and fares).

3.2.2 Weather

All data concerning weather conditions are provided by the nearest climate stations, under the management of INTA (National Institute of Agricultural Technology). There are three monitoring stations in the study area that report climate data every fifteen minutes. Several spectra of variables are available, but we use those of temperature, wind and precipitation so that we can compare our outcomes with those of Arana et al. (2014).

3.2.3 Economics

We use monthly wage index historical series (base 2012=100) and the economic activity monthly estimator from *Instituto Nacional de Estadísticas y Censos* (INDEC). As the price index during some periods is not fully reliable, we use the consumer price index from the statistic institute of Buenos Aires city. From the same source we also obtained the automotive fleet evolution of the city. From *Confederación de Entidades del Comercio de Hidrocarburos y Afines de la República Argentina* (CECHA), we got the daily petrol final price. Finally, we use an online source to get exchange rate series[1].

3.3 Methodology

3.3.1 Basic Statistics

First of all, we evaluate if passenger data is stationary, i.e. if the statistical properties do not change over time (mean, variance, etc.). In other words, we would expect series without trends, with constant variance and no seasonality or autocorrelation (weak stationarity process definition). Several statistical forecasting methods are based on this assumption (e.g., ARIMA models). Without this property, models could provide meaningless sample statistics and invalid estimations.

In Figure 3.6 we decompose[2] the passenger series to evaluate how it behaves. It does not seem to exhibit a long-run trend effect and decomposed residuals also seem quite random. However, the series exhibits a strong cyclical pattern. This seasonality is apparently due to the weekdays-weekends effect. Every weekend, there is a massive fall in passengers, as you can appreciate in Figure 3.7, where we zoom the first weeks of the series.

[1] source:investing.com

[2] we apply a multiplicative decomposition expressed as: $y = x * trend * seasonality * error$

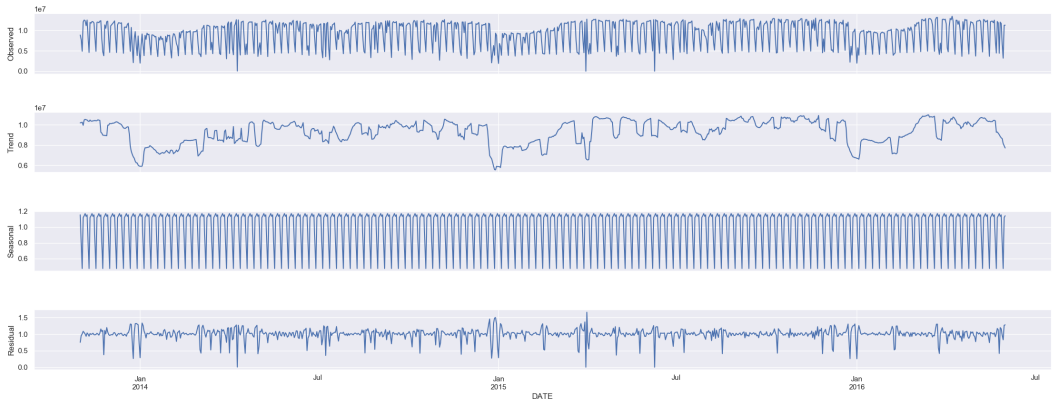


Figure 3.6. Multiplicative Decomposition

Figure 3.7. Seasonality Pattern

One popular stationarity test is the Augmented Dickey-Fuller test –ADF- (Fuller, 1976; Dickey and Fuller, 1979). Under the null hypothesis the time-series is a unit root, i.e. not only the shocks have permanent effects but also the variance is time dependent. If the null hypothesis is not rejected, we should apply difference operators to the series.

Despite the ADF rejected the null hypothesis (i.e., the series does not follow a unit root process), we take the seven difference of the series to smooth the seasonality effect.

As can be appreciated in Figure 3.8, the passenger series looks more stable and it seems to follow a stationarity process. There are still some strong peaks which are actually related to general strikes (specifically, 10th April 2014, 31st March 2015, 9th June 2015). The results of the ADF test are reported in Table 3.1.

We also check stationarity on the predictors to avoid spurious regressions, and we differentiate if it is needed (Newbold and Granger, 1974).

Additionally, we run a Darwin-Watson Test –DW- (Durbin and Watson; 1950, 1951, 1971) to check autocorrelation. The test statistic is approximately equal to $2*(1-r)$ where r is the sample autocorrelation of the residuals. Thus, for $r=0$, indicating no serial correlation, the test statistic equals 2. We get a DW of 1.872, therefore, we do not find evidence of serial correlation.

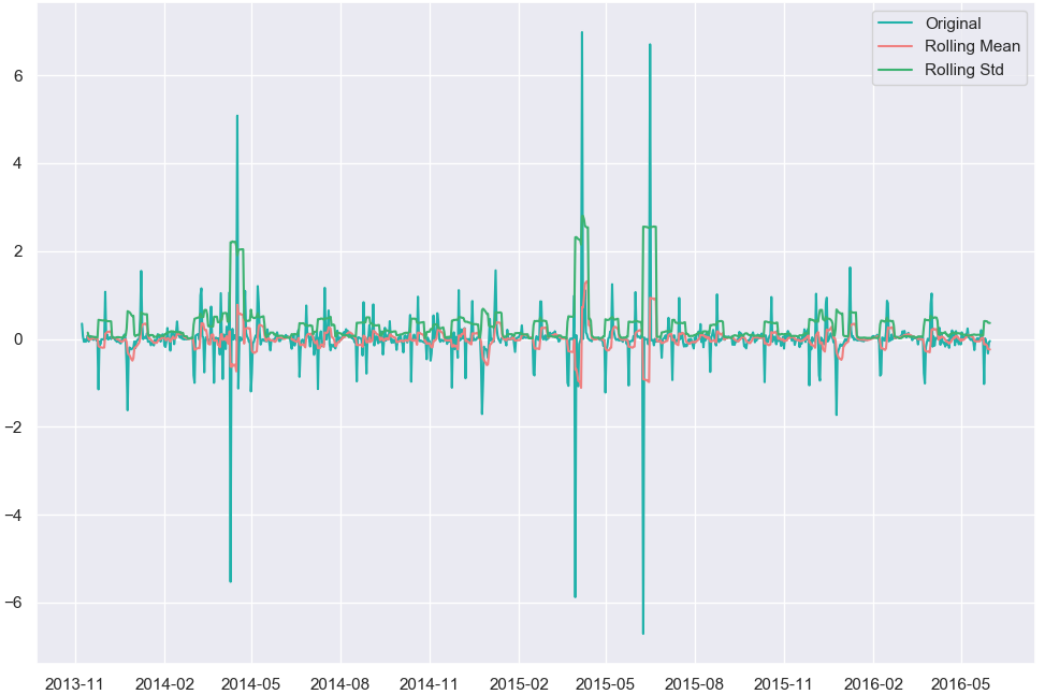


Figure 3.8. Series Stationarity

$Passengers_t - Passengers_{t-7}$	
Test Statistic	-14.45
p-value	0.00
Lags Used	7
Number of Observations Used	928
Critical Value (1%)	-3.44
Critical Value (5%)	-2.86
Critical Value (10%)	-2.57

Table 3.1. Augmented Dickey-Fuller Test

3.3.2 Ordinary Least Squares

Our first model is a log linear Ordinary Least Squares (OLS). The initial model can be stated as:

$$log(y_t) = \beta x_t + \epsilon$$

Where y_t is the seven difference of number of passengers in the period t, and x is the vector that represents explanatory variables.

To complete the comparative analysis, we take as our starting point the predictive study reported by Arana et al. (2014), which undertake a multiple linear regression analysis, using smart card and weather data, to predict demand. We also introduce a number of additional considerations which enriches our estimations considerably. The results of the OLS regression are shown in Table 3.2.

	coef	**std err**	**t**	$P > \|t\|$	**[0.025**	**0.975]**
National Day	-0.9238	0.114	-8.070	0.000	-1.149	-0.699
Workday	-0.0930	0.056	-1.666	0.096	-0.203	0.017
Strike	-3.0036	0.345	-8.706	0.000	-3.681	-2.326
Government	-0.0000	0.000	-1.454	0.146	-0.000	0.000
Precipitation (mm)	-0.0045	0.002	-2.544	0.011	-0.008	-0.001
Temperature (C°)	0.0129	0.009	1.516	0.130	-0.004	0.030
Strong Wind	0.2765	0.252	1.098	0.273	-0.218	0.771
Exchange Rate (ARS/USD)	0.0457	0.557	0.082	0.935	-1.048	1.140
Economic Activity Index	-0.0138	0.016	-0.881	0.378	-0.045	0.017
Real Wage Index	-9.4126	6.108	-1.541	0.124	-21.405	2.580
Relative Fare (Metro/Bus)	2.5289	0.907	2.787	0.005	0.747	4.311
Oil Final Price (ARS)	-0.9386	0.534	-1.758	0.079	-1.987	0.110
Relative Fare (Train/Bus)	0.2102	0.093	2.248	0.025	0.027	0.394
Fleet	0.0004	0.000	6.376	0.000	0.000	0.000
Nominal Fare (ARS)	-0.1180	0.101	-1.172	0.241	-0.316	0.080
Trend	0.0001	0.000	0.914	0.361	-0.000	0.000
Vehicle Fleet	0.0000	0.000	0.961	0.337	-0.000	0.000

Dep. Variable:	Pax - Pax(-7)	**Model:**	OLS
No. Observations:	701	**R-squared:**	0.301
Durbin-Watson:	1.872	**Adj. R-squared:**	0.285

Table 3.2. OLS Regression Results

We proceeded to include several time variables (Hauer, 1971), which we deem important for quantifying demand and which are essential to ensure a good quality analysis. We added a dummy variable controlling a government transition (*Government*) and a dummy variable related to general strikes (*Strike*). We incorporated

a variable that differentiates working days from non-working days -i.e., Saturdays, Sundays- (*Workday*) and a variable for national holidays (*National Day*).

We then added the three weather variables considered by Arana et al. (2014) with slight differences. We included as a dummy the notion of *Strong Wind*, that is, a wind with a velocity greater than 50 km/hour on the Beaufort scale, as we consider it to be more readily interpretable in terms of its explanatory power. Also we introduce the precipitation level (*Precipitation (mm)*) and the average temperature (*Temperature (C^o)*).

We complemented the above with the only offer relevant variable, namely *Fleet*, which computes the total number of buses circulating each day. All the other characteristics (e.g., quality, new routes, etc.) are static and, therefore, we assume there is no endogeneity between supply and demand.

The last variables we introduced are related to the bus fare, prices and the economic situation. Following Holmgren (2007) we added a real wage index to take into account the income-elasticity (*Real Wage Index*), the price of petrol (*Oil Final Price (ARS)*) and the relative price between bus and substitutes (*Relative Fare (Metro/Bus)* and *Relative Fare (Train/Bus)*). Finally, we included a variable related to the economic activity in CABA (*Economic Activity Index*), the nominal fare (*Nominal Fare (ARS)*) and a trend variable.

All the variables present their expected signs, however not all of them are significant. In summary, during national days and strikes, less passengers are recorded. Precipitation is the only weather variable which seems to have a significant impact. The results also show that bus is a substitute service for train and metro. Additionally, the size of the fleet presents a positive relation with a larger number of travelers and there is no evidence of income-elasticity effects. Finally, the nominal fare does not show any significant impact.

The results of this OLS time-series model could however be very influenced by serial correlation, i.e. the error term observations could be correlated. A pattern in the error term (when we are assuming that it is white noise) could bias the significance of the explanatory variables.

As many time-series models work under the assumption that residuals are white noise (i.e., a random sequence which cannot be predicted), we should evaluate if they are stationary and if they do not present any serial correlation problem. For testing stationarity, we apply the ADF test which rejects the null hypothesis of unit root (see Table 3.3).

Now we know that the errors are stationary, we evaluate serial correlation. We apply the Ljung-Box test (Ljung and Box, 1978) which checks if whether any of a group of autocorrelations of the series are statistically different from zero (under the null hypothesis the data is independently distributed). It tests the overall ran-

Residuals	
Test Statistic	-11.39
p-value	0.00
Lags Used	7
Number of Observations Used	693
Critical Value (1%)	-3.44
Critical Value (5%)	-2.86
Critical Value (10%)	-2.57

Table 3.3. OLS: Augmented Dickey-Fuller Test

domness based on a number of lags. We use the rule proposed by Hyndman and Athanasopoulos (2018) and set the number of lags equal to $min(2m, T/5)$, where m is the period of seasonality and T is the length of the time series.

	L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**L9**	**L10**	**L11**	**L12**	**L13**	**L14**
QLJB	3.78	4.32	4.67	4.80	4.85	12.50	73.68	74.51	74.79	75.65	75.88	76.17	76.55	76.89
pvalue	0.05	0.12	0.20	0.31	0.43	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.4. OLS: Serial Correlation

As shown in Table 3.4, we reject the null hypothesis of no serial correlation for several lags and we cannot consequently affirm that residuals are white noise. Before continuing with the machine learning algorithms, we are therefore going to introduce a SARIMAX model.

3.3.3 SARIMAX

Autoregressive Integrated Moving Average (ARIMA) is a time-series model which uses two polynomials terms to describe stationary stochastic processes: autoregressors (AR) and moving averages (MA), and an initial differentiating step. Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of ARIMA models which explicitly takes into account seasonal components: $ARIMA(p, d, q)$ $x(P, D, Q)_s$ where P is the number of seasonal autoregressive (SAR) terms, D is the number of seasonal differences, and Q the number of seasonal moving average (SMA). If we also add exogenous input terms we get the SARIMAX definition.

For choosing the correct model specification we have to evaluate the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The ACF describes the autocorrelation between observations and observations at a prior time step. The PACF is the autocorrelation between observations in a time series with observations at prior time steps with the relationships of intervening observations removed. When autocorrelation is present, the error term follows a pattern, invali-

dating essential assumptions of time-series models. If autocorrelation exists, there is some information that could be explaining the movements of the dependent variable and we are not able to capture. In Figure 3.9 we plot the ACF and the PACF.

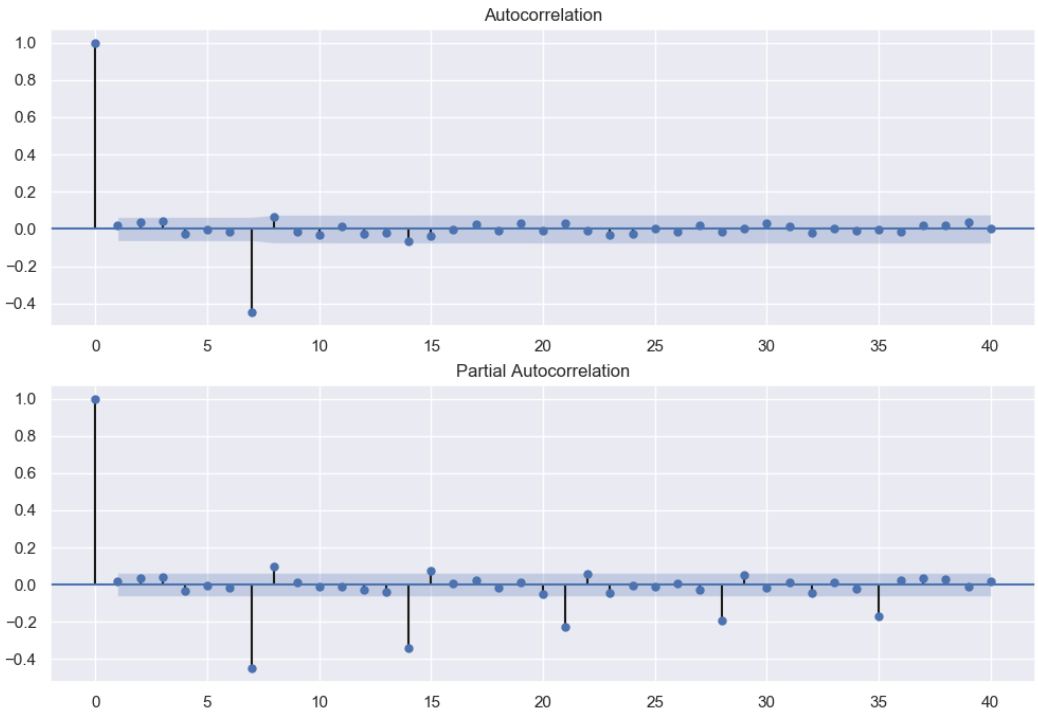


Figure 3.9. Series Stationarity

The passenger variable does not seem to show a strong autocorrelation with days before. Remember that now the dependent variable is the difference between the number of passengers and the number of passengers one week before. With this in mind, there is some correlation with the same difference but one week before. We therefore add an autoregressive variable of order 7 (AR7) to the predictors. Besides, in the partial autocorrelation plot, we can appreciate a seasonality pattern every 7 lags. This can be solved by using seasonality ARIMA models, particularly, by applying moving averages terms for the cyclical pattern (SMA).

	Statistic	**p-value**
Ljung-Box (Q)	24.28	0.98

Table 3.5. SARIMAX: Serial Correlation

Table 3.5 reports the Ljung-Box test after we added to the model the temporal terms. As can be appreciated, we cannot now reject the null hypothesis of no serial

correlation. Knowing that the residuals are white noise, we therefore present the results of the SARIMAX model (see Table 3.6)[3].

	coef	std err	z	$P > \|z\|$	[0.025	0.975]
National Day	-0.9210	0.123	-7.506	0.000	-1.162	-0.681
Strike	-3.3073	0.075	-43.891	0.000	-3.455	-3.160
Precipitation (mm)	-0.0048	0.002	-2.155	0.031	-0.009	-0.000
Relative Fare (Metro/Bus)	1.9968	0.877	2.278	0.023	0.279	3.715
Relative Fare (Train/Bus)	0.1625	0.059	2.774	0.006	0.048	0.277
Fleet	0.0003	0.000	8.084	0.000	0.000	0.000
Nominal Fare (ARS)	-0.0865	0.044	-1.984	0.047	-0.172	-0.001
AR7	-0.4725	0.015	-32.131	0.000	-0.501	-0.444
ma.S.L7	0.0789	0.020	3.889	0.000	0.039	0.119
ma.S.L14	-0.2104	0.025	-8.277	0.000	-0.260	-0.161
Dep. Variable:	Pax - Pax(-7)	**Model:**	SARIMAX(1, 0, 2, 7)			
No. Observations:	701	**R-squared:**	0.532			
Ljung-Box (Q):	24.28	**Prob(Q):**	0.98			

Table 3.6. SARIMAX Model Results

In comparison with the OLS model, we obtain quite similar results. But now, either AR7 and the seasonal MA are very significant.

However, this tells us nothing in terms of demand forecasts; indeed, causality would appear not to be directly relevant. Clearly, we would all expect the number of passengers to decrease on national days and for rain to act as a deterrent to mobility. What bus operators and transport regulators need also to know is the likely number of passengers at any specific time, that is, they need accurate forecasts. In other words, they need answers to such questions as: How many passengers will there be if it rains tomorrow? The smart card is an excellent tool – a highly innovative technology – that provides us with daily feedback for use in predictive analyses. Smart cards facilitate the construction of an unparalleled base learner that is constantly improving itself. Given the availability of these data, the next step is to start using algorithms that can exploit this advantage.

Therefore, we are going to compare traditional models in three ways: interpretability, predictive power in out-of-sample data, and demand-elasticity estimation. But before, let us briefly introduce some of the most popular supervised machine learning algorithms.

[3]It is important to remark that we followes the practical approach proposed by Andrews et al. (2013) to select the appropriate SARIMAX model.

3.3.4 Machine Learning Algorithms

Penalized Linear Regressions

Penalized linear regression (PLR) is a method designed to overcome some of the problems associated with OLS, basically that of overfitting (i.e., the impossibility to generalize well from the training data to out-of-sample data). They allow degrees of freedom to be reduced to fit data and model complexity. They are especially good methods when degrees of freedom are tight.

We adopt a shrinkage approach in our estimations, that is, we augment the error criteria that is being minimized with a data-independent penalty term (or regularization parameter α). The problem we seek to solve by using α is a problem inherent to all predictions, namely, the trade-off between bias and variance (or overfitting). $P(\beta)$ is the penalty function that can take several forms. The most common are the ridge regression (Tikhonov and Arsenin, 1977)-which uses the squares of β- and the lasso regression (Santosa and Symes, 1986)-which uses the absolute values of β. Elastic net regression combines these two methods, as shown below:

$$\beta_0^*, \beta^* = argmin_{\beta_0^*, \beta^*}(\frac{1}{m}\sum_{i=1}^{m}(y_i - (\beta_0 + x_i\beta))^2 + \lambda\alpha_L|\beta| + (1-\lambda)\alpha_R\beta^T\beta)$$

So that when $\lambda = 1$, it corresponds to a lasso penalization, and when $\lambda = 0$, it corresponds to a ridge regression.

In summary, here, what we seek is a good quality trade-off between bias and variance. If we can achieve a greater reduction in variance than the corresponding increase in bias, then we can obtain higher accuracy. To choose α we applied the Elastic net regression with iterative fitting along a regularization path (using 5-Folds cross validation). The α that minimizes the mean-squared error is 0.21, as can be seen in the Figure 3.10.

Binary Decision Trees

A decision tree consists of a set of tree-structured decisions that takes Boolean decisions. A set of features is tested from the root node, and by a recursive process (which optimizes the splits) we obtain the prediction when the leaf node is reached. The disadvantages of this kind of model are that it tends to be noisy and to overfit (Last et al., 2002; Quinlan, 2014). To avoid these problems, we present a variety of ensemble methods. Ensemble methods use a set of algorithms that combine different predictions (base learners like binary decision trees) and the combination of these results offers better outcomes than those obtained from random guessing.

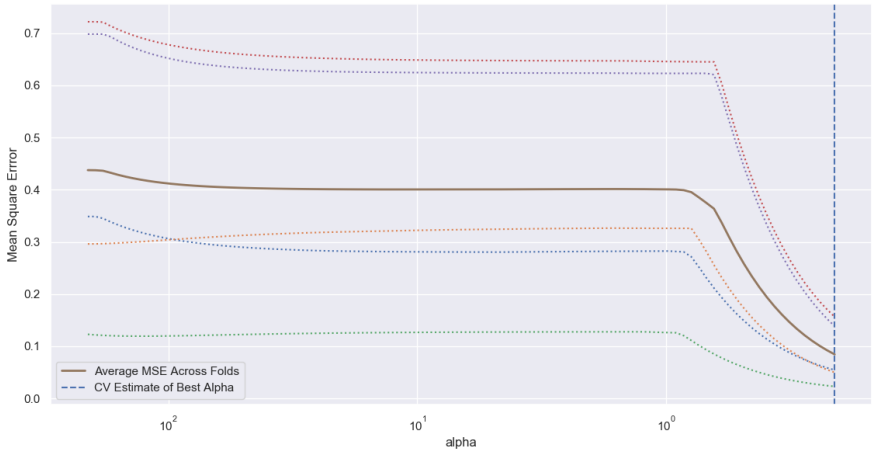


Figure 3.10. Elastic Net 5-Fold CV

Aggregation Algorithm (Bagging)

Bagging (Breiman, 1996) uses bootstrap samples (i.e., samples with repetition) from the training data and then it trains a base-learner in each of these samples. The combination of independent base learners leads to a decrease in the loss function. Finally, it takes a simple average of their outcomes.

Bagging primarily addresses the variance error, but it has some issues with the bias error. This means that it needs good depth (given that it is a simple model that generates splitting points concentrated in the same place).

Random Forest

Bagging only constructs trees using bootstrap samples of data, whereas random forest –RF-(Breiman, 2001) also uses a random sample on predictors before each node is split, until the tree conditions are fulfilled. This ensures greater independence between trees, because of the combination of bootstrap samples and random draws of predictors. Consequently, we can take advantage of averaging a large number of trees (and so obtain better levels of variance reduction). Similarly, we can gain in terms of bias reduction, because we can employ a very large number of predictors (more even than the number of observations), and local feature predictors can play a role in tree construction.

In conclusion, this method has all the advantages of bagging combined with a lower propensity to overfit (each tree fits, or overfits, a part of the training set, and in the end these errors cancel each other out, at least partially), and, as we see below, it is easier to tune than gradient boosting (GB). Thus, RF works particularly well with

fully grown decision trees (low bias, high variance). Moreover, it tackles the error reduction task in the opposite way, that is, by reducing variance. In contrast to GB, trees are made uncorrelated to maximize the decrease in variance, but RF cannot reduce bias (which is slightly higher than that of a simple binary tree). Hence, the need for large unpruned trees, so that the bias is as low as possible at the outset.

In contrast to decision trees, we can reduce the number of observations in the terminal nodes, because RF is less likely to overfit. However, we maintain the same splits rule. Finally, we have to set the number of predictors sampled, and this is a key tuning parameter that will affect performance. There exist several rules, but the most common is $k = log_2(n) + 1$ (as recommended by Breiman, 2001).

Gradient Boosting

What GB (Friedman, 2001) does is to train a set of trees, where every tree is trained on the error of the previous ensemble models. GB starts in the same way as bagging, but it focuses on the areas that present most mistakes. This gives a better approximation, without the need for greater depth, which is an essential advantage. In contrast with RF, GB works well when based on weak learners in terms of high bias and low variance (even as small as decision stumps). GB reduces the error primarily by reducing bias, and to some extent the variance, by aggregating the output from many models.

In summary, RF trains with a random sample of data in addition to randomizing features. It relies on this randomization to give a better generalization performance on out-of-sample data. GB, on the other hand, additionally seeks to find the optimal linear combination of trees, where the final model is the weighted sum of predictions of individual trees applied to the training data.

3.4 Results

We evaluate the results in three different ways: interpretability, predictive power and demand elasticity. In doing so, we split our data-set in three sequential samples (see Figure 3.11):

1. Test (54 obs): This is the period with the last fare increase (approximately 80%) and which we are going to use at the end of the study to test our hypothesis about demand elasticity (from April 2016 to June 2016).

2. Train (713 obs): 80% of the total sample without taking into account the test sample (from November 2013 to October 2015). During this period we have

two fare increases: On 1st of January 2014 fares increased 67% and on 1st of July 2014 fares increased 16%).

3. Valid (177 obs): 20% of the total sample without taking into account the test sample (from October 2015 to April 2016).

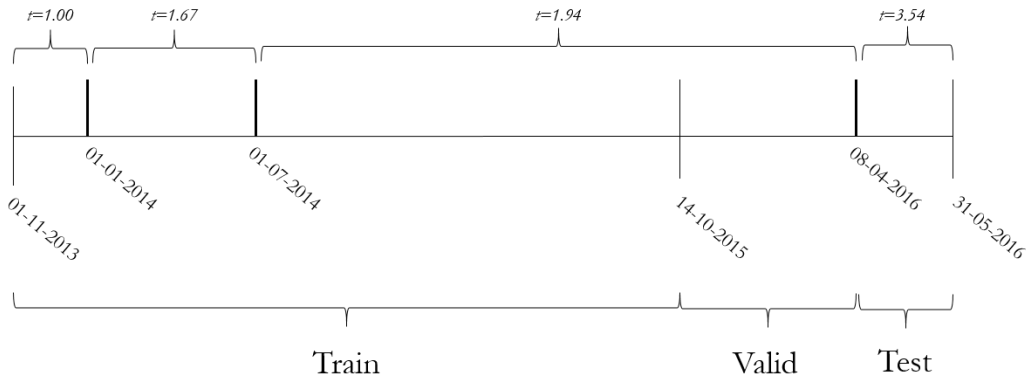


Figure 3.11. Sample Split. *Note: Fare evolution is given by t with base November 2013 = 1.00*

3.4.1 Interpretability

The main advantage of linear models is interpretability, they are simple and intuitive. As their pursuit unbiased estimator we can get not only the relevant variables but also their magnitudes (by partial derivatives). In contrast, machine-learning algorithms tends to be "black box" (Klaiber and von Haefen, 2011; Zhao et al., 2018). As their purse out-sample predictive power, the variance reduction is obtained by biasing the estimators. However, we can develop feature importance ranking (Vellido et al., 2012; Molnar, 2018) to get the significant variables of the model. Feature importance provides a score which measures the importance of each attribute to improve the performance (in terms of reducing the loss function). It is the average improvement of each attribute in every decision tree (when a splitting point is selected) weighted by the number of observations in the nodes.

In Table 3.7, we provide a comparison of the techniques presented earlier. In general, they show quite similar results. The main problem with feature importance ranking is however that we do not know the direction and the magnitude of the effects. Nevertheless, we can see that all the models agree in the importance of the seasonal variables $AR7$, $ma.S.L7$ and $ma.S.L14$ (ML algorithms selected them as the most relevant variables). Then, *National Day* is considered one of the most

Variables	OLS	SARIMAX	Elastic Net	Binary Tree	Bagging	Random Forest	Gradient Boosting
National Day	-0.9238*** (0.114)	-0.921*** (0.123)	-	-	0.05	0.22	0.05
Workday	-0.093* (0.056)	-	-	-	0.00	0.00	-
Strike	-3.0036*** (0.345)	-3.3073*** (0.075)	-	-	-	-	-
Government	-0.0000 (0.000)	-	-	-	-	-	-
Precipitation (mm)	-0.0045** (0.002)	-0.0048** (0.002)	- 0.00	0.00	0.00	0.02	0.00
Temperature (°C)	0.0129 (0.009)	-	-	0.00	0.00	0.03	0.00
Strong Wind	0.2765 (0.252)	-	-	-	-	-	-
Exchange Rate (ARS/USD)	0.0457 (0.557)	-	-	-	0.00	0.01	0.00
Economic Activity Index	-0.0138 (0.016)	-	-	-	-	0.00	-
Real Wage Index	-9.4126 (6.108)	-	-	-	0.00	0.02	0.00
Relative Fare (Metro/Bus)	2.5289*** (0.907)	1.9968** (0.877)	-	-	0.00	0.02	-
Oil Final Price (ARS)	-0.9386* (0.534)	-	-	-	-	0.00	-
Relative Fare (Train/Bus)	0.2102** (0.093)	0.1625*** (0.059)	-	0.01	0.00	0.09	0.00
Fleet	0.0004*** (0.000)	0.0003*** (0.000)	0.00	-	0.00	0.31	0.00
Nominal Fare (ARS)	-0.118 (0.101)	-0.0865** (0.044)	-	-	-	0.00	-
Trend	0.0001 (0.000)	-	0.00	0.00	0.00	0.01	0.00
Vehicle Fleet	0.00004248 (0.000)	-	- 0.00	-	-	0.00	-
AR7	-	-0.4725*** (0.015)	-	0.00	0.10	0.45	0.01
ma.S.L7	-	0.0789*** (0.020)	0.25	0.67	0.13	0.70	0.65
ma.S.L14	-	-0.2104*** (0.025)	0.38	1.00	1.00	1.00	1.00

Table 3.7. Interpretability: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$. *Standard Errors are given in parentheses*

significant variables (only Elastic Net and Binary Tree do not consider it). *Precipitation (mm)*, *Fleet* and relative fares are also pointed out by some methods. Surprisingly, strikes are not considered by any machine learning algorithm, something that could be explained by the minimum sample restrictions in the nodes splits.

3.4.2 Predictive Power

As explained earlier, the main advantage of ML algorithms is the ability to reduce the variance, i.e., to get the best out-sample predictive power. We use two well-known comparative metrics to evaluate this aspect: 1) Mean Squared Error, 2) Mean Absolute Error.

Mean Squared Error

Mean Squared Error (MSE) calculates the squared difference between predicted values ($\hat{y}$) and the actual values y.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Absolute Error

Mean Absolute Error (MAE) is the average over the absolute differences between prediction and actual values in the test sample. The main difference is that it is less sensitive to outliers compared to MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

A comparison of our main results (see Table 3.8) shows that supervised machine learning methods provide persistently much better results in terms of error performance than linear models (RF gets a reduction of 93-100% compared to the OLS model). This means we obtain much better accuracy in the out-sample data, which is achieved by eliminating the linearity restrictions imposed by traditional methods and by exploiting the potential of tree-based models.

3.4.3 Demand Elasticity

If we focus on the SARIMAX model, in terms of income elasticity, there is no clear relation between income and passengers. The problem with income-elasticity is that income could be correlated with other variables, which might make the effect noisy.

Model	**MSE**		**MAE**	
OLS	0.14	-	0.21	-
SARIMAX	0.43	207%	0.35	101%
Elastic Net	0.04	-69%	0.14	-49%
Binary Decision Tree	0.04	-68%	0.08	-88%
Bagging	0.02	-87%	0.09	-87%
Random Forest	**0.01**	**-93%**	**0.07**	**-100%**
Gradient Boosting	0.03	-76%	0.10	-75%

Table 3.8. Predictive Power

For instance, higher income will raise the public transport demand demand. But, for car owners, higher income not increase their demand. Since the probability of owning a car increases with income, this might affect the total impact (Holmgren, 2007).

With regard to other transport modes, there is evidence of substitution effect between train, metro and bus. In contrast, we did not find evidence that supports relation between bus and petrol price or automotive fleet.

However, what we are mainly concerned is about own price elasticity. Holmgren (2007) collected 81 articles and estimated price-elasticities ranging from -0.009 to -1.32, with a mean value of -0.38 (i.e., on average, public transport has an inelastic demand). However, by region, the price elasticity was -0.75 in Europe and -0.59 in USA, Canada and Australia.

In the city of Buenos Aires case and during the period analyzed, there were three bus fare increases (see Table 3.9).

Period	ARS		USD	
jan-13	1.81	-	0.37	-
jan-14	3.02	67%	0.46	24%
jul-14	3.49	16%	0.43	-7%
apr-16	6.38	83%	0.44	2%

Table 3.9. Fare Evolution

At the same time, Argentina have suffered from high inflation rates (an accumulative inflation rate of 134%). As a result, real fares have barely changed. If there is any effect on demand, it should therefore be a money illusion effect (controlling for the other factors). Despite the fact that we have controlled by real income, price of competitors (train price, metro price and petrol price) and other variables, all the economic variables have a very similar trend evolution (see Figure 3.12).

At the outset, we set aside the period from 8th April 2016 to 31st May 2016.

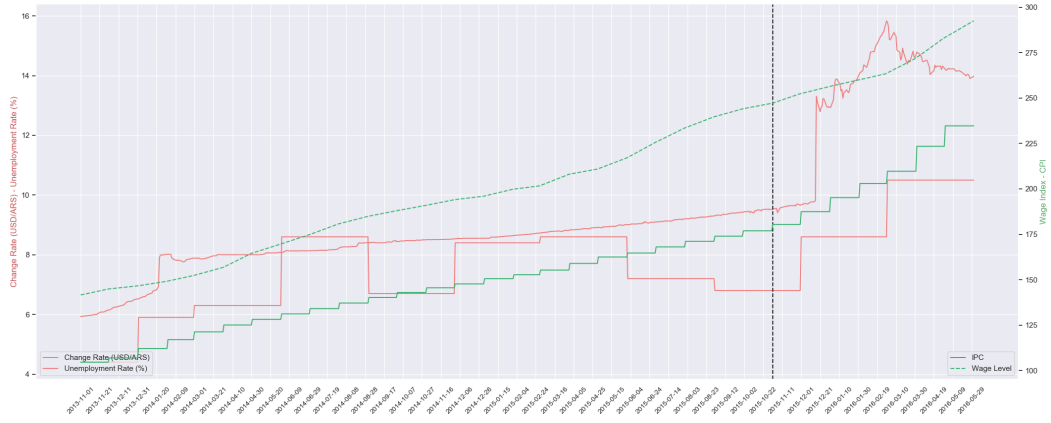


Figure 3.12. Economic Variables Evolution

During this period, fares have increased around 80%. Those months give us a great opportunity to evaluate nominal price elasticities.

If we accept the results of our models, only SARIMAX model pointed out *Nominal Fare* as a significant variable. However, as can be seen in Figure 3.13, comparing days before and after the last increase, it seems to be a reduction in the absolute number of passengers during the firsts weeks (a possible money illusion effect).

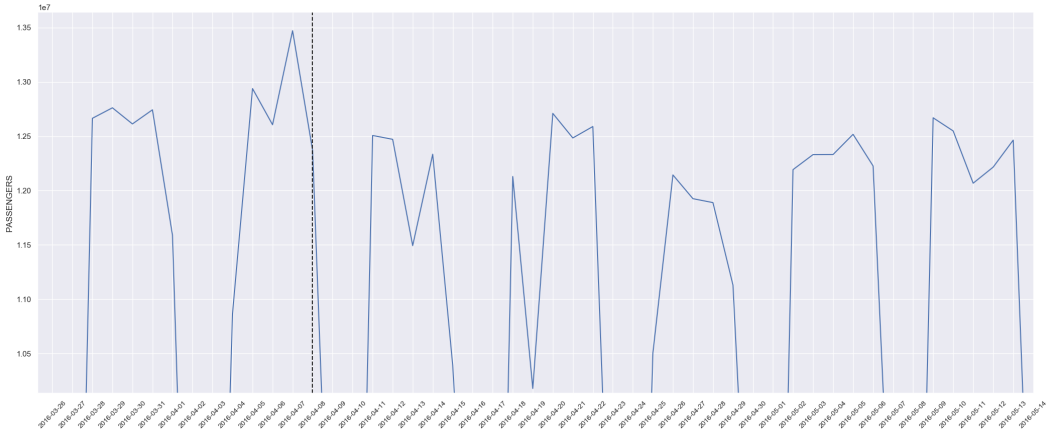


Figure 3.13. Passenger (in millions) Before and After the Last Fare Increase

This simple plot analysis reveals that fare increase might affected transport demand in the short-term. Although it was a money illusion effect (because real fare was almost the same), our hypothesis is that during the first days, passengers negatively reacted to the new fares trying to reduce their consume of public transport (or the number of interconnections). However, as days went by, passengers internalized the effect and started to use the public transport as usual. In the overall effect, we will not see any change, because only a few days were affected and the algorithms may confuse them with sparse or noisy data. But, if we are able to decompose the

weekly effect, we would expect to find a significant pattern during the first days and none effect during the weeks after.

In terms of prediction, as seen in Figure 3.14, every model seems to be capable to forecast the passenger's trend.

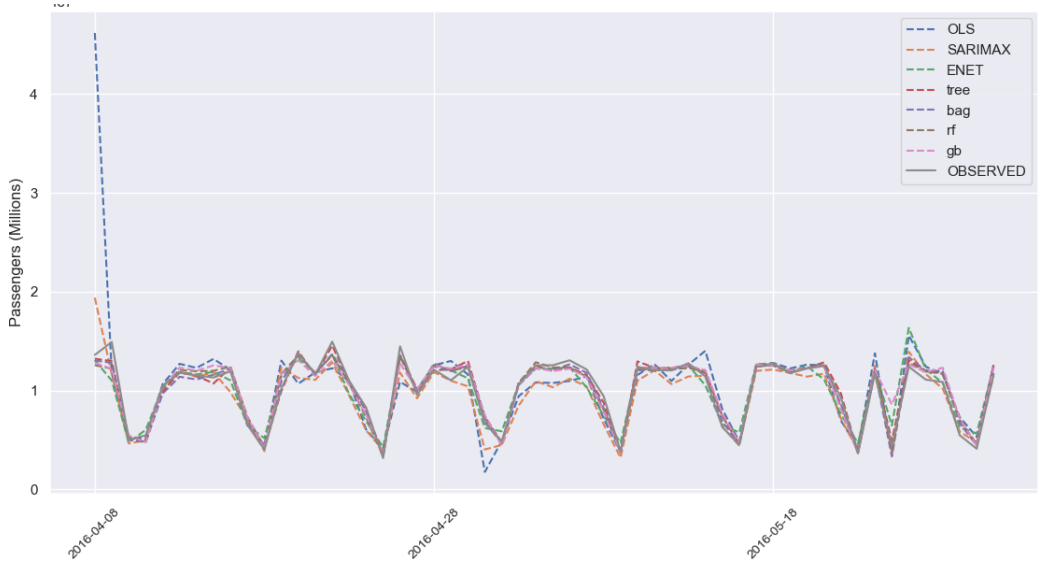


Figure 3.14. Test Prediction

A problem with machine learning algorithms is that they are not able to use partial derivative to get elasticity effects (because the coefficients are biased). However, we can apply sensitivity analysis like the arc elasticity to compute a mean average effect (Nunes et al., 2016; Alsger et al., 2016; Jung and Sohn, 2017; Miller, 2018; Zhao et al., 2018). Essentially, we can measure how the predictions respond to changes in the input fare.

We therefore propose to compute the arc elasticity as the difference between the observed and the predicted values, assuming there was no fare increase, formally,

$$E(Y) = \frac{y(t') - \hat{y}(t)}{t' - t} * t'/y$$

However, it persists an error between the true and the predicted values (as showed in Predictive Power subsection). Consequently, we propose to correct this estimation by adding the error difference between the observed values ($y(t')$) and the predicted values with the fare increase ($\hat{y}(t')$). In the best of cases, this error tends to zero. Introducing this correction term, we assure that the results are not biased because of prediction errors. We can therefore calculate the arc elasticity as:

$$E(Y) = \frac{y(t') - \hat{y}(t) - e}{t' - t} * t'/y$$

where $e = y(t') - \hat{y}(t')$.

We take for example the RF model, which performed better in terms of out-sample prediction (see Figure 3.15). Both, *PREDICTED (t')* (or $\hat{y}(t')$) and *PREDICTED (t)* (or $\hat{y}(t)$) seem to show no difference, i.e., the model is not affected by fare changes. Consequently, they are mutually canceled: the differences between *OBSERVED(t')* and *PREDICTED(t)* can be only assigned to the prediction error $y(t') - \hat{y}(t')$. Thereby, the final result in the elasticity equation will be $E(Y) = 0$, which is perfectly consistent with unresponsiveness to fare changes.

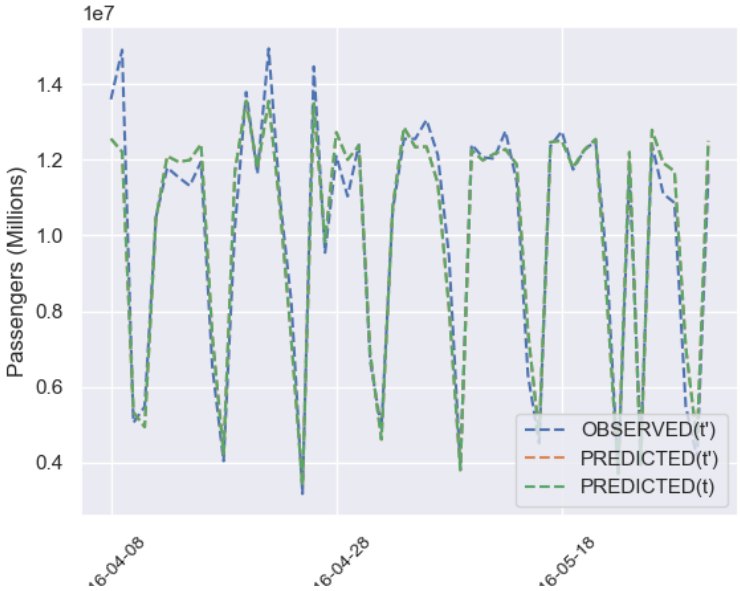


Figure 3.15. Test Prediction with Random Forest

Finally, we apply a difference in means test to evaluate if the mean difference between the observed values and the predicted values with no fare increases (minus the error term) is significant. Under the null hypothesis, two independent samples have identical average values. Table 3.10 reports the results.

Model	**Elasticity**	**p-value**
SARIMAX	-0.3095**	0.047
Elastic Net	-	1.000
Binary Decision Tree	-	1.000
Bagging	-	1.000
Random Forest	-	1.000
Gradient Boosting	-	1.000

Table 3.10. Elasticity and Difference in Means Test

None of the ML models can reject the null hypothesis of difference in means and demand elasticity remains in zero. On the contrary, the linear model estimates a demand elasticity of -0.309, similar to the -0.30 often mentioned as rule of thumb in public transport demand elasticity (Goodwin, 1992; Oum, et al., 1992, Bresson, et al., 2003; Holmgren, 2007).

What is more interesting is however to examine what happened day by day. We therefore expand the elasticity of the linear model using an accumulative mean (see Figure 3.16).

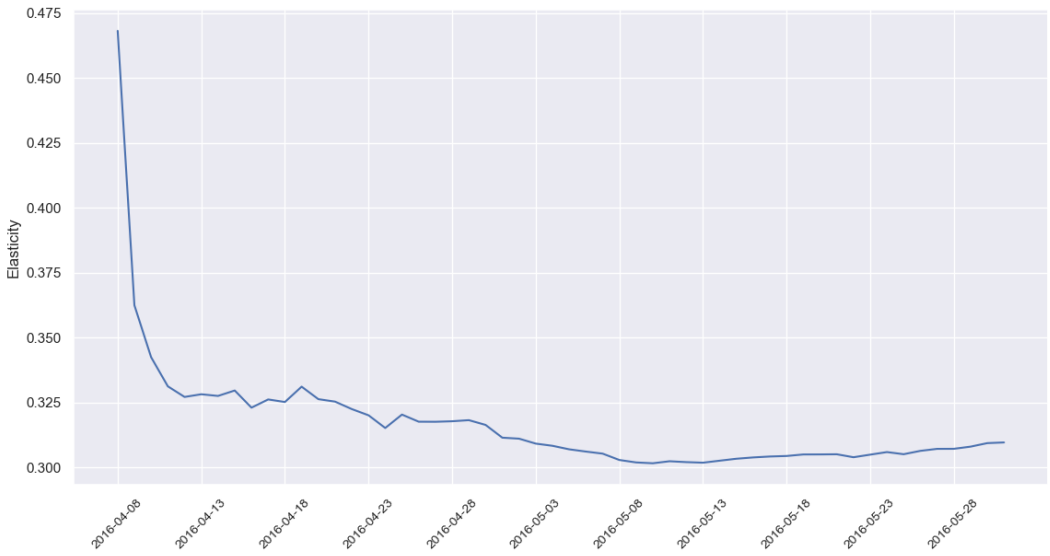


Figure 3.16. SARIMAX: Expanded Elasticity

SARIMAX model shows an initial impact immediately after the fare increase (with absolute elasticity higher than 0.45). Then, as days go by, the elasticity tends to rapidly decrease until it reaches a stable behavior. This is consistent with the hypothesis we have stated before: During the first days, passengers negatively reacted to a nominal effect, but then, they tended to readjust their consume level to a slightly lower level than before.

3.5 Conclusions

While previous studies have exploited smart cards to predict demand, they have typically adopted an unbiased orientation to address a problem that is clearly predictive in nature. Here, we take an error focus and propose different supervised machine learning algorithms for application to the smart card data obtained from the SUBE system employed in Argentina. Specifically, we have examined the bus system operating in the Autonomous City of Buenos Aires (CABA), thanks to the

rich and reliable source of information it provides. We have compared a set of machine learning algorithms with traditional time-series regression models in an effort to identify the method that provides the best result in terms of: prediction, interpretability and demand elasticity.

First, we have shown how machine learning algorithms, perform better (in terms of predictive power) than linear demand predictions for public transport services. Despite the concerns expressed in previous studies about out-of-sample data (which means it is unclear how well they perform on new data), we conclude that supervised machine learning methods, in general, perform 49-100% better than traditional unbiased methods.

Second, while machine learning algorithms are often associated with "black-box" results, we conclude that in terms of interpretability they show very similar findings as linear models. We have undertaken a feature analysis to determine which variables have most impact on demand predictions and we find that the variables with the greatest impact are not those directly related to weather conditions (except precipitations), but rather that time variables are persistently the most influential (seasonal terms, national days, strikes). Other relevant factors are cross elasticities with other public transport services and fleet size. But we do not find evidence about elasticities with respect to price of petrol, income or vehicle fleet.

Finally, we applied a sensitivity analysis to measure demand elasticity. In doing so, we proposed a corrected arc elasticity formulation to control for prediction error. We have evaluated this formulation during a period where nominal fares increased around 80%. None of SML algorithms showed responsiveness to change in nominal fares. However, SARIMAX model revealed a short-term demand elasticity pattern of 0.31, with an initial shock higher than 0.45. Then, as days go by, the elasticity tends to rapidly decrease until it reaches a stable behavior.

4 Abnormal Pattern Prediction in the Insurance Market: Fraudulent Property Claims

4.1 Introduction[1]

Predicting abnormalities in environments with highly unbalanced samples and a huge mass of unlabeled data is receiving more attention as new technologies are developed (e.g., time-series monitoring, medical conditions, intrusion detection, detecting patters in images, etc.). A typical example of such a situation is provided by fraud detection (Hodge, 2004, Weiss, 2004, Phua et al., 2010, Ahuja and Singh, 2017). In general, we only have partial information about fraud cases, as well as possibly some information about false positives, that is, cases that are considered suspicious but which prove to be cases of non-fraud. The problem here is that we cannot label these cases as "non-fraud" simply because they were initially considered suspicious. For this reason, we know nothing about non-fraud cases. Moreover, fraud tends to be an outlier problem, given that we are dealing with atypical values with respect to regular data. Hence, it is likely that we only dispose of information about an extremely small sample. Yet, it so transpires, that this information is extremely useful and should not be discarded. In contrast we have a considerable amount of data that may contain fraud and or non-fraud cases and, as such, we cannot treat these data using traditional supervised algorithms.

To represent this typical case we apply an innovative semi-supervised methodology to a real fraud case. Specifically, we draw on information provided by a leading insurance company as we seek to predict fraudulent insurance claims[2]. In general terms, fraud insurance claims fall into two categories: one, those that provide only partial or untruthful information in the policy contract; and, two, those that are based on misleading or untruthful circumstances (including exaggerations). It has

[1] Article published at Data Science Journal. Reference: Palacio, S.M., 2019. Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning. Data Science Journal, 18(1), p.35. DOI: http://doi.org/10.5334/dsj-2019-035

[2] The study is part of the development of a fraud detection system that was implemented in 2018.

been estimated that cases of detected and undetected fraud represent up to 10% of all claims in Europe (The Impact of Insurance Fraud, 2013), accounting for around 10-19% of the payout bill.

In the sector, the main services contracted are automobile and property insurance, representing 76% of total claim costs. However, while many studies have examined automobile fraud detection (see, for example, Artís et al., 1999 and 2002; Belhadji et al., 2000; Stefano and Gisella, 2001; Brockett et al., 2002; Phua et al., 2004; Viaene et al., 2007; Wilson, 2009; Nian et al., 2016), property fraud has been largely neglected, perhaps because detection is more difficult as witnesses are infrequent or they are typically cohabitants. One representative case is Bentley (2000) who uses fuzzy logic rules to detect suspicious property insurance claims in an unbalanced dataset of 98 fraudulent claims and 20,000 unknown cases. They got accuracy rates of 60% based on three artificial assumptions of 0%-5%-10% proportions of suspicious cases in the unknown claims.

In addition, private companies rarely share real fraud datasets and keep this information private to not reveal competitive details. Very small number of studies have therefore been implemented as fraud systems in insurance companies (few examples are Major and Riedinger, 1992; Cox, 1995).

Our main objective is therefore to present a variety of semi-supervised machine learning models applied to a fraud insurance detection problem. In so doing, we aim to develop a methodology capable of improving results in classification anomaly problems of this type. The key being to avoid making assumptions about the unknown fraud cases when resolving reoccurring practical problems (skewed data, unlabeled data, dynamic and changing patterns) since this can bias results.

Our reasoning for using semi-supervised models is best explained as follows. First, as pointed out by Phua et al. (2010), **skewed data** is a challenge in many fraud studies. They find that more than 80% of the papers analyzed have a percentage of fraud cases below 30%. For instance Bentley's (2000) study have only 0.5% fraud cases whilst 99.5% are unknown, and Foster and Stine (2004) use just 2,244 cases of bankruptcies compared to 2.9 million credit card transactions to predict personal bankruptcy. Statistically speaking, fraud can be considered a case of outliers, that is, points in the data-set that differ significantly from the remaining data. Outliers do not mean noise. We refer to outliers as observations that remarkably deviate from normal data. Fraud is typically classified as abnormal behavior or a sudden change of patterns and therefore differs from noise (Barnett and Lewis, 1994, Hodge and Austin, 2004; Weiss, 2004; Aggarwal, 2015). Thus, skewed and unlabeled data is a natural consequence. Such anomalies often result from unusual events that generate anomalous patterns of activity. Were we to use unsupervised models – that is, were we to assume that we are unable to distinguish between fraudulent

and non-fraudulent cases – what we defined as outliers, noise or normal data would be subjective and we would have to represent that noise as a boundary between normal data and true anomalies without any information. But, as mentioned, the number of fraud cases detected is small; however, they constitute a useful source of information that cannot be discarded.

Second, supervised models are inappropriate because, in general, we face a major problem of claim misclassifications when dealing with fraud detection (Artís et al., 2002) which could generate a substantial mass of **unknown data**. Fraud detection, typically, comprises two stages: first, it has to be determined whether the claim is suspicious or not (Viaene et al., 2007); and, second, all cases considered suspicious have to be examined by fraud investigators to determine whether the claim is fraudulent or not. This means that unsuspicious cases are never examined, which is reasonable in terms of efficiency, especially if the process cannot be automatized. Insurance adjusters have little time to perform an exhaustive investigation. Yet, the process does provide us with partial information, that is, labels for what is a small sample. Clearly, using a supervised model in this instance adds bias to the confusion matrix. Essentially, we will detect severe bias in false negatives and, therefore, many cases which are in fact fraudulent will be predicted as being non-fraudulent (Phua et al., 2004). Indeed, when using supervised algorithms we assume that the system in place is capable of discerning perfectly between fraudulent and non-fraudulent claims, an outcome that in practice is infrequent and referred to in the literature as an "omission error" (Bollinger and David, 1997; Poterba and Summers, 1995).

Finally, when fraud investigators analyze claims, they base their analysis on a small suspicious subset from previous experience and tend to compare cases to what they consider to be "normal" transactions. As data volume and the velocity of operative processes increases exponentially, human analysis becomes poorly adapted to **changing patterns** (Lei and Ghorbani, 2012).

Clearly, the information provided in relation to cases considered suspicious is more likely to be specified correctly once we have passed the first stage in the fraud detection process. This information will be useful for a part of the distribution (i.e., it will reveal if a fraudulent claim has been submitted), which is why it is very important this information be taken into account. For this reason, fraud detection in insurance claims can be considered a semi-supervised problem because the ground truth labeling of the data is partially known. Not many studies have used hybrids of supervised/unsupervised models. Williams and Huang (1997) cluster data from a Medicare Insurance, treating each cluster as a class and use them to construct a decision tree that generate decision rules. As a result, they are able to identify possible groups of interest for further investigation. Williams (1999) continues down

the same line, using a system that is able to evolve with the progression of claims. Brockett et al. (1998) study automobile bodily injury insurance claims in over 387 cases. They ask loss-adjusters and investigators to group the cases by level of suspiciousness, and later use Self Organizing Maps to cluster the data and re-label it. However, basing the construction of clusters on subjective boundaries between fraud and non-fraud can bias the outcomes.

Other semi-supervised models use normal observable data to define abnormal behavioral patterns: Aleskerov et al. (1997) use past behavior as normal data to predict anomalies using Neural Networks. Kokkinaki (1997) detects atypical transactions based on users' profiles normal behavior. Murad and Pinkas (1999) identify fraudulent patterns in phone-calls finding "significant deviation" from the normal data (which is based on profiling). Kim et al. (2003) use normal product sales to detect anomalous sales patterns. However, these studies assume we have information about normal behavior, which is not always the case, and, it is questionable whether or not the normal observable data was correctly defined as normal in the first place.

We therefore seek to make three contributions to the literature: First, we apply semi-supervised techniques to an anomaly detection problem while trying to solve three combined problems: skewed data, unlabeled data and change in patterns, **without making any subjective assumption** that can bias the results. Second, we create a metric based on the logic behind the F-Score which permit us to evaluate the purity of abnormalities in the clusters. Finally, we build a fraud detection system which is applied to an actual property insurance claim fraud problem, using a real-world data-set provided by a leading insurance company.

4.2 Data

We use an insurance fraud data-set provided by a leading insurance company in Spain, initially for the period 2015-2016. After sanitization, our main sample consists of 303,166 property claims, some of which have been analyzed as possible cases of fraud by the Investigation Office (IO)[3].

Of the cases analyzed by the IO, 48% proved to be fraudulent. A total of 2,641 cases were resolved as true positives (0.8% of total claims) during the period under study. This means we do not know which class the remaining 99.2% of cases belong to. However, the fraud cases detected provide very powerful information, as they reveal the way in which fraudulent claims behave. Essentially, they serve as the pivotal cluster for separating normal from abnormal data.

[3]The system applied before to detect fraud corresponds to a rule based methodology.

A data lake was constructed during the process to generate sanitized data. A data lake is a repository of stored raw data, which includes structured and unstructured data in addition to transformed data used to perform tasks such as visualizing, analyzing, etc. From the data lake, we obtain 20 bottles containing different types of information related to claims. A bottle is a subset of transformed data which comes from an extract-transform-load (ETL) process preparing data for analysis. These bottles contain variables derived from the company's daily operations, which are transformed in several aspects. In total we have almost 1,300 variables. We briefly present them in Table 4.1 to help explain which concepts were included in the model.

Bottles	Descriptions
ID	ID about claims, policy, person, etc.
CUSTOMER	Policyholder's attributes embodied in insurance policies: name, sex, age, address, etc.
CUSTOMER_PROPERTY	Customer related with the property data.
DATES	Dates of about claims, policy, visits, etc.
GUARANTEES	Coverage and guarantees of the subscribed policy.
ASSISTANCE	Call center claim assistance.
PROPERTY	Data related to the insured object.
PAYMENTS	Policy payments made by the insured.
POLICY	Policy contract data, including changes, duration, etc.
LOSS ADJUSTER	Information about the process of the investigation but also about the loss adjuster.
CLAIM	Brief, partial information about the claim, including date and location.
INTERMEDIARY	Information about the policies' intermediaries.
CUSTOMER_OBJECT_RESERVE	The coverage and guarantees involved in the claim.
HISTORICAL_CLAIM	Historical movements associated with the reference claim.
HISTORICAL_POLICY	Historical movements associated with the reference policy (the policy involved in the claim).
HISTORICAL_OTHER_POLICIES	Historical movements of any other policy (property or otherwise) related to the reference policy.
HISTORICAL_OTHER_CLAIM	Historical claim associated with the reference policy (excluding the claim analyzed).
HISTORICAL_OTHER_POL_CLAIM	Other claim associated with other policies not in the reference policy (but related to the customer).
BLACK_LIST	Every participant involved in a fraudulent claim (insured, loss-adjuster, intermediary, other professionals, etc.)
CROSS VARIABLES	Several variables constructed with the interaction between the bottles.

Table 4.1. The 20 Data Bottles and their descriptions extracted from a Data Lake created for this particular case study.

4.3 Methodology

If we have labeled data, the easiest way to proceed is to separate regular from outlier observations by employing a supervised algorithm. However, in the case of fraud, this implies that we know everything about the two classes of observation, i.e., we would know exactly who did and did not commit fraud, a situation that is extremely rare. In contrast, if we know nothing about the labeling, that is, we do not know who did and did not commit fraud, several unsupervised methods of outlier detection can be employed, e.g. isolation forest (Liu et al., 2008), one-class support vector machines (Schölkopf et al., 2001; Manevitz and Yousef, 2001) and elliptic envelopment (Rousseeuw and Driessen, 1999). However, they tend to be less precise and we have to assume some subjective boundary.

If, however, we have some label data about each class, we can implement a semi-supervised algorithm, such as label propagation (Zhu and Ghahramani, 2002) or

label spreading (Zhou et al., 2004). Yet, these methods require that we have some information about every class in our problem, something that is not always possible. Indeed, disposing of label data information about each class is quite infrequent in certain practical problems. Additionally, we face the problem of unbalanced data, which means we rarely have clean and regular data representing the population. In fraud problems, as a norm, the data is highly imbalanced, which results in a high but biased success rate.

In the light of these issues, we propose a semi-supervised technique that can assess not only a skewed data-set problem but also one for which we have no information about certain classes. In this regard, fraud detection represents an outlier problem for which we can usually identify some, but not all, of the cases. We might, for example, have information about false positives, i.e., investigated cases that proved not to be fraudulent. However, simply because they have raised suspicions mean they cannot be considered representative of non-fraudulent cases. In short, what we usually have are some cases of fraud and a large volume of unknown cases (among which it is highly likely cases of fraud are lurking).

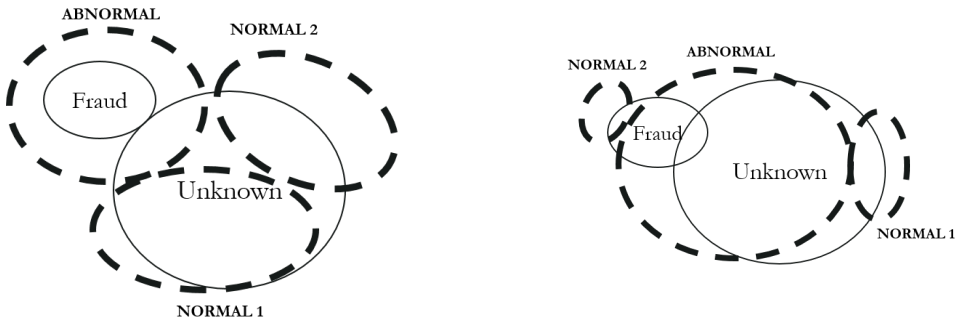
Bearing this in mind, we propose the application of unsupervised models so as to relabel the target variable. To do this, we use a new metric that measures how well we approximate the minority class. We can then transform the model to a semi-supervised algorithm. On completion of the relabeling process, our problem can be simplified to a supervised model. This allows us not only to set an objective boundary but to obtain a gain in accuracy when using partial information, as Trivedi et al. (2015) have demonstrated.

4.3.1 Unsupervised Model Selection

We start with a data-set of 303,166 cases. The original data was collected for business purposes, therefore a lot of time was put into sanitizing the data-set. It is important to remark that we set aside a 10% random subset for final evaluation. Hence, our data-set consists of 270,479 non-identified cases and 2,370 cases of fraud.

The main problem we face in this unsupervised model is having to define a subjective boundary. We have partial information about fraud cases, but have to determine an acceptable threshold at which an unknown case can be considered fraudulent. When calculating unsupervised classification models, we reduce the dimensions to clusters. Almost every algorithm will return several clusters containing mixed-type data (fraud and unknown). Intuitively, we would want the fraud points revealed to be highly concentrated into just a few clusters. Likewise, we would expect some non-revealed cases to be included with them, as in Figure 4.1a. On the other hand, we would want to avoid situations in which abnormal and normal cases

are uniformly distributed between groups, as in Figure 4.1b. Thus, a limit of some kind has to be defined. But, how many of the "unknown" cases can we accept as being fraudulent?



(a) Desired Clusters (b) Non-desired Clusters

Figure 4.1. Possible clusters. Figure 4.1a shows a separable and compact cluster of the abnormal points. On the other side, Figure 4.1b shows abnormal and normal cases uniformly distributed.

A boundary line might easily be drawn so that we accept only cases of detected fraud or we accept every possible case as fraudulent. Yet, we know this to be unrealistic. If we seek to operate between these two extremes, intuition tells us that we need to stay closer to the lower threshold, accepting only cases of fraud and very few more, as Figure 4.2 illustrates.

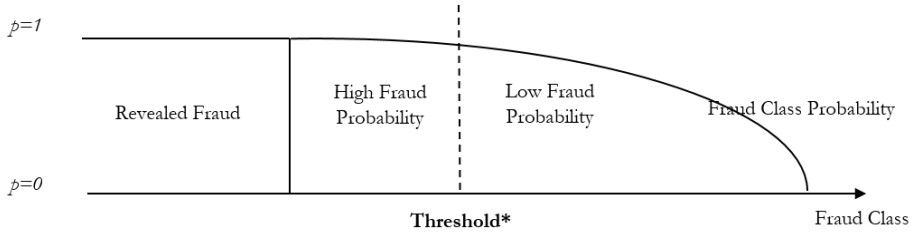


Figure 4.2. Schematic representation of the desired threshold, which is expected to split high fraud probability cases from low fraud probability cases.

But once more, we do not know exactly what the correct limit is. In this way, however, we have created an experimental metric that can help us assign a score and, subsequently, define the threshold. This metric, which we shall refer to as the cluster score (CS), calculates the weighted homogeneity of clusters based on the minority and majority classes.

$$CS_\alpha = (1+\alpha^2)\frac{C1*C2}{C1+C2*\alpha^2} \qquad \text{with } \alpha > 0,\ \alpha \in \mathbf{R}$$

Essentially, it assigns a score to both the minority-class (C1) and the majority-class (C2) clusters based on the weighted conditional probability of each point. The CS expression clearly resembles the well-known F-Score[4], which is a measure of the test's accuracy. Particularly, C1 and recall, and C2 and precision, pursue the same objectives, which in our case is to capture the maximum amount of fraud cases while also paying attention to the quality of those cases. The CS measure permits us to maximize homogeneity in the clusters. Since C1 and C2 are part of the same subset space, we have to make trade-offs (just as with recall and precision) between the optimization of C1 homogeneity and C2 homogeneity.

Moreover the α parameter allows us to maximize the homogeneity we are more concerned about. If for example, we want to obtain a more homogeneous C1 (the fraud cluster would include almost every possible case of revealed fraud), we can set a higher α, taking into account that it possibly makes the C2 homogeneity worse.

C1 Score

Suppose an unsupervised model generates J clusters: $\{C^1, C^2, ..., C^J\}$. The number of cases in cluster C^j is denoted by n^j.

The C1 score calculates the probability that a revealed (i.e., confirmed) fraud case belongs to cluster C^j and this probability is weighted by the total number n^j_{fraud} of fraud cases in that cluster C^j, divided by the total number of N_{fraud} of revealed fraud cases in the dataset.

$$C1 \; = \; \frac{\sum_{j=1}^{J} \frac{n^j_{fraud}}{n^j} * n^j_{fraud}}{N_{fraud}} \; \in [0,1]$$

Basically, we calculate the fraction of fraud cases in each cluster j (n^j_{fraud}/n^j) and we weight these fractions by the corresponding number of fraud cases in cluster j (n^j_{fraud}).

Our objective is to maximize C1. This means ensuring all revealed fraud cases are in the same clusters. The limit C1=1 implies that all J clusters only contain revealed fraud cases. Therefore, we have to balance this function with another function.

C2 Score

C2 is the counterpart of C1. The C2 score calculates the probability that an "unknown" case belongs to cluster C^j and this probability is weighted by the total

[4] F-Score is defined as $F_\beta \; = \; (1+\beta^2)\frac{precision*recall}{recall+\beta^2*precision}$.

number ($n^j_{unknown}$) of unknown cases in that cluster C^j, divided by the total number of unknown cases in the data-set ($N_{unknown}$):

$$C2 \;=\; \frac{\sum_{j=1}^{J} \frac{n^j_{unknown}}{n^j} * n^j_{unknown}}{N_{unknown}} \;\in [0,1]$$

Notice that $n^j_{fraud} + n^j_{unknown} = n^j$. The objective is the same as that above in the case of C1: to cluster the class of unknown cases without assigning revealed fraud cases to these clusters.

Cluster Score

Individually maximizing C1 and C2 leaves us in an unwanted situation. Basically, they are both trying to be split. Therefore, when we maximize one, we minimize the other. If we maximize both together, this results in a trade-off between C1 and C2, a trade-off in which we can choose. Moreover, as pointed out above, we actually want to maximize C1 subject to C2. Consequently, the fraud score is constructed as follows:

$$CS_\alpha \;=\; (1+\alpha^2)\frac{C1 * C2}{C1 + C2 * \alpha^2} \qquad \text{with } \alpha > 0,\ \alpha \in \mathbf{R}$$

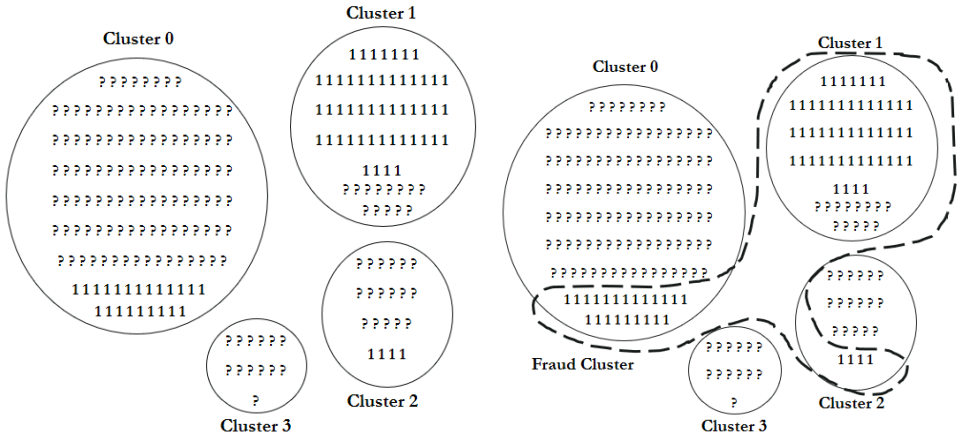
If $\alpha = 1$, C1 and C2 will have the same weight. But if we assign $\alpha > 1$, this will reduce the weight of C2 (if $\alpha < 1$, this will reduce the weight of C1). It is important to highlight that the actual function of the cluster score is to choose between algorithms (based on the purity of the cluster construction) and α is the way to balance C1 and C2.

In conclusion, with this CS we have an objective parameter to tune the unsupervised model because it permits us to homogeneously evaluate not only different algorithms but also their parameters. While it is true that there exists a variety of internal validation indices, this metric differs in that it can also exploit information about the revealed fraud cases. That is, we take advantage of the sample that is labeled fraud to choose the best algorithm, something that internal validation indices are not able to accomplish. The only decision that remains for us is to determine the relevance of α. A numerical example can be consulted in Appendix 1.

We should stress that each time we retrieve more information about the one-class cases that have been revealed, this threshold improves. This is precisely where the entropy process of machine learning appears. As fraud is a dynamic process that changes patterns over the time, using this approach the algorithm is capable of adapting to those changes. In the one-class fraud problem discussed above, we

start with an unknown distribution for which some data points are known (i.e., the fraud sample). Our algorithms, using the proposed CS metric, will gradually get closer to the best model that can fit these cases of fraud, while maintaining a margin for undiscovered cases. Now, if we obtain new information about fraud cases, our algorithms will readjust to provide the maximum CS again. As the algorithms work with notions based on density and distances, they change their shapes to regularize this new information.

Once the best unsupervised model is attained (i.e., the model that reaches the maximum CS), we need to decide what to do with the clusters generated. Basically, we need to determine which clusters comprise fraudulent and which comprise non-fraudulent cases. The difficulty is that several clusters will be of mixed-type: e.g., minority-class points (fraud cases) and unidentified cases, as in Figure 4.3a, where the 0s are unidentified cases and the 1s are minority-class points.



(a) Unsupervised output (b) Relabeled clusters

Figure 4.3. Cluster Example Output. Figure 4.3a shows an example of a cluster algorithm output over a sample of data points. Figure 4.3b shows how the Cluster Score chooses the points that are relabeled as fraud cases (points inside the doted line).

In defining a threshold for a fraud case, we make our strongest assumption. Here, we assume that if a cluster is made up of more than 50% of fraud cases, this cluster is a *fraud cluster*, otherwise, it is a *non-fraud cluster*. The distinction introduced is clear: The non-fraud cluster is no longer an unidentified cluster. By introducing this assumption, we state that they are actually non-fraudulent cases. This definition acts as the key for our transition into a semi-supervised model. The assumption may seem unrealistic but, as we will see later, the best unsupervised models are capable of generating clusters with a proportion greater than 95% of fraud cases. We can, therefore be even more stringent with this assumption.

As Figure 4.3b shows, cluster 1, being composed of more than 50% fraud cases, now forms part of the more general fraud cluster, together, obviously, with the fraud cases already detected. The remaining cases that do not belong to such a dense fraud cluster are now considered non-fraud cases.

As mentioned, before applying the unsupervised algorithm, we had to make a huge effort to sanitize the original data since it was collected for business purposes. This included: handling categorized data, transforming variables, bad imputation, filtering, etc. at each bottle level. Finally, we transformed the 20 bottles at a claim level and put them together in a unique table which formed our model's input.

After that, before using this data as input, we made some important transformations. First, we filled the missing values given that many models are unable to work with them. There are simple ways to solve this, like using the mean or the median value of the distribution. Since we did not want to modify the original distribution, we implemented a multi-output Random Forest regressor (Breiman, 2001), to predict the missing values based on the other columns. The idea was, for each column that had missing values, we used the column as a target variable. We trained with the part without missing values, and by using the other features, we predicted the target variable.

We iterated this process in every column that had missing values (0.058% of the total values were missing). We also measured the performance of this technique using the R-squared, which is based on the residual sum of squares. Our R-squared was 89%.

Second, we normalized the data to, later, be able to apply a Principal Component Analysis (PCA), and also because many machine-learning algorithms are sensible to scale effects. Those using Euclidean distance are particularly sensitive to high variation in the magnitudes of the features. In this case, we used a robust scale approach[5] that is less affected by outliers since it uses the median value and the interpercentile ranges (we chose 90%-10%). In general, standard normalization is a widely use method. However, as in this case we are paying special attention to outliers, a mean approach might not be the best option. Outliers can often influence the sample mean/variance in a negative way. The robust scale approach removes the median and scales the data according to a quantile range. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Median and interquartile range are then stored to be used on test/new data.

Third, we applied Principal Component Analysis to resolve the high dimensionality problem (we had almost 1,300 variables). This method reduces confusion in

[5]We use the formulation $Z = (x - x_{median})/(p_{90} - p_{10})$

the algorithms and solves any possible collinearity problems. PCA decomposes the data-set in a set of successive orthogonal components that explain a maximum amount of the data-set's variance. When setting a data-set's variance threshold, a trade-off between over-fitting and getting the variation in the data-set is made. We chose a threshold of 95% (recommended threshold is between 95% and 99%), which resulted in 324 components. After this transformations, the unsupervised algorithm can thus be summarized as seen in Algorithm 1.

The main reason is that it has a low noise sensitivity as it ignores small variations in the background (based on a maximum variation basis). While it is true that there are several non-linear formulations for dimensionality reduction that may get better results, some studies have actually found that non-linear techniques are often not capable of outperforming PCA. For instance Van Der Maaten et al. (2009) compared PCA versus twelve non-linear dimensionality reduction techniques on several data-sets and they couldn't conclude that the non-linear techniques outperformed PCA.

4.3.2 Supervised Model Selection

We now have a redefined target variable that we can continue working with by applying an easy-to-handle supervised model. The first step involves re-sampling the fraud class to avoid unbalanced sample problems. Omitting this step, means that our model could be affected by the distribution of classes, the reason being that classifiers are in general more prone to detect the majority class rather than the minority class. We, therefore, oversample the data-set to obtain a 50/50 balanced sample. We use two oversampling methods, Adaptive Synthetic Sampling Approach (ADASYN) by He et al. (2008), and balanced subsampling. ADASYN finds the n-nearest neighbors in the minority class for each of the samples in the class. It creates random samples from the connections and adds a random small value to the points in order to scatter them and break the linear correlation with the parent point. The balanced subsample method on the other hand, does not need to create synthetic points since the samples used are already balanced. The balanced samples are obtained by using weights inversely proportional to class frequencies for each iteration in a supervised tree based algorithm.

The second step, involves conducting a grid search and a Stratified 5-fold cross-validation (CV) based on the F-Score[6] to obtain the optimal parameters for three different models: extreme randomized tree -ERT- (Geurts et al., 2006), gradient boosting -GB- (Freund and Schapire, 1996) and a light XGB -LXGB- (Ke et al., 2017). Cross-validation is a great way to avoid over-fitting, i.e., failing to predict new data. We train using $k-1$ folds (data subsets) and we validate our model by

[6]The F-Score was constructed using $\beta = 2$, as we needed to place greater weight on the recall.

Algorithm 1: Unsupervised algorithm

Data: Load transformed data-set. Oversample the fraud cases in order to have the same amount as the number of unknown cases.

for $k \in K = \{model_1, model_2, ...\}$*where* K *is a set of unsupervised models.* **do**

  for $i \in I$ *where* I *is a matrix of parameter vectors containing all possible combinations of the parameters in model* k **do**

    We fit the model k with the parameters i to the oversampled data-set.;

    We get the J clusters: $\{C^1, C^2, ..., C^J\}$ for the combination $\{k, i\}$, i.e., $C_{k,i} = \{C^1_{k,i}, C^2_{k,i}, ..., C^J_{k,i}\}$;

    For $C_{k,i}$ we calculate C1 Score and C2 Score and we obtain the cluster score $CS_{k,i}$, based on the acceptance threshold t^*.;

    Save the cluster score result $CS_{k,i} \in CS_{K,I}$, where $CS_{K,I}$ is the cluster score vector for each pair $\{k, i\}$.;

  end

end

Choose the optimal CS^* where $CS^* = max\{CS_{K,I}\}$

Relabel the fraud variable using the optimal clustering model derived from CS^*. Each unknown case in a fraud cluster is now equal to 1, known fraud cases are equal to 1 and remaining cases are equal to 0.

testing it on the remaining fold. To prevent an imbalance problem in the folds, we use a stratified k-folds strategy which returns subsets containing approximately the same distribution of classes as the original data-set.

We have to be careful not to over-fit the model during the cross-validation process, particularly when using oversampling methods. Step one and step two, therefore have to be executed simultaneously. Oversampling before cross-validating would generate samples that are based on the total data-set. Consequently, for each $k-1$ training fold, we would include very similar instances in the remaining test fold, and vice versa. This is resolved by first, stratifying the data, and then oversampling the $k-1$ folds, without taking into account the validation fold. Finally, we concatenate all the predictions.

Additionally, we combine the supervised models using stacking models. Stacking models is combining different classifiers, applied to the same data-set, and getting different predictions that can be "stacked" up to produce one final prediction model. The idea is very similar to k-fold cross validation, dividing the training set into several subsets or folds. For all $k-1$ folds, predictions are obtained by using all the supervised models (called the base models). The predictions are stored to be used as features for the stacking model in the full training data-set. Finally, a new model (the stacking model or the Meta model) is fitted to the improved data-set. The stacking model can discern whether a model performs well or poorly, which is very useful since one model might have high performance when predicting fraud, but not when predicting non-fraud, and vice versa. The combination of both could therefore improve the results. We try three different ways of combining classifiers, modifying the Meta model: GB and LXGB with Meta ERT, GB and ERT with Meta LXGB, and LXGB and ERT with Meta GB.

Once we have the optimal parameters for each model, we calculate the optimal threshold that defines the probability of a case being fraudulent or non-fraudulent, respectively.

Finally, we identify the two models that perform best on the data-set – the best acting as our main model implementation, the other controlling that the predicted claims are generally consistent. The algorithm can be summarized as seen in Algorithm 2.

Algorithm 2: Supervised algorithm

Data: Load relabeled data-set.

for $model_i \in M' = \{M, S\}$ *where* M *is the set of supervised individual models* M *and* S *the set of stacking models from* M **do**

  for *{*$train_k$*,* $test_k$*} folds in the Stratified k-Folds* **do**

    if *Oversample==True* **then** $train'_k = oversample(train_k)$ where oversampling is applied to 50/50 using the ADASYN method.;

    else $train'_k = train'_k$ and the balanced subsampling option is activated.;

    Fit the $model_i$ in $train'_k$, where $model_i \in M' = \{M, S\}$.;

    Get predicted probabilities p_k of $test_k$ using $model_i$.;

    Save the probabilities p_k in P_i, where P_i is the concatenation of $model_i$'s probabilities.

  end

  for $\forall t_i \in [0, 1]$*, where* t *is a probability threshold of the* $model_i$ *to consider a case as fraudulent* **do**

    if $P_i \geq t_i$ **then** $P_i = 1$;

    else $P_i = 0$;

    Using P_i, where now P_i is a binary list, we calculate, $FScore_{i,t} = (1 + \beta^2) * \frac{precision * recall}{recall + \beta^2 * precision}$ with $\beta = 2$.;

    Save $FScore_{i,t}$ in $FScore_i$, a list of vectors of $model_i$ with $FScore$ results for each t.

  end

  We get $FScore_i^* = max\{FScore_i(t)\}$.

end

4.4 Results

4.4.1 Performance

Table 4.2 shows the main unsupervised modeling results of the tuning process. We tried different combinations of distance based models, density based models and outlier models: Mini-Batch K-Means (Sculley, 2010), Isolation Forest (Liu, 2008), DBSCAN (Ester et al., 1996), Gaussian Mixture and Bayesian Mixture (Figueiredo and Jain, 2002). Mini-batch K-Means is not only much faster than the other models, it also provides the best results. It is similar to K-Means++, both using the Euclidean distance between points as the objective function, however it can also reduce computation time. Subsets of the input data are taken and randomly sampled in each iteration, converging more quickly to a local solution.

Model	**n Clusters**	**C1**	**C2**	**CS** ($\alpha = 2$)
Mini-Batch K-Means	4	96.6%	96.6%	96.6%
Isolation Forest	2	51.5%	51.1%	51.4%
DBSCAN	2	50.2%	49.8%	50.1%
Gaussian Mixture	5	95.0%	95.0%	96.3%
Bayesian Mixture	6	96.5%	96.4%	96.5%

Table 4.2. Unsupervised model results

C1 indicates that the minority-class (fraud) clusters comprise approximately 96.59% of minority data points on a weighted average. In contrast, C2 indicates they are made up of 96.59% of unknown cases. As can be seen in Table 4.3, more than 95% of the cases in the central cluster are fraudulent (well above our 50% assumed threshold), but it also contains an additional 6,047 unknown cases (Cluster 0 now contains an additional 5,890 cases, and Cluster 1 an additional 157 cases). This is our core fraud cluster and the one we use when renaming the original labels.

After relabeling the target variable (with the Mini-Batch K-Means output), we calculate the supervised models performance using Stratified 5-Fold CV on the dataset. The results of each of the supervised models and of the stacking models is shown in Table 4.4.

As can be appreciated, we have two recall values. The cluster recall is the metric derived when using the relabeling target variable. The original recall emerges when we recover the prior labeling (1 if it was fraud, 0 otherwise). As can be seen, the results are strikingly consistent. We are able to predict fraud cluster with a recall of up to 89-97% in every case. But, more impressively yet, we can capture the original fraud cases with a recall close to 98%. The precision is slightly lower, but in almost

Clusters	Fraud	Percentage
0	0	2%
0	1	98%
1	0	99%
1	1	1%
2	0	100%
2	1	0%
3	0	1%
3	1	99%

Table 4.3. Oversampled Unsupervised Mini-Batch K-Means

all cases it is higher than 67%. These are particularly good results for a problem that began as an unsupervised high-dimensional problem with an extremely unbalanced data-set.

Model	Cluster Recall	Original Recall	Precision	F-Score
ERT-ss	0.9734	0.9840	0.6718	0.8932
ERT-os	0.9647	0.9819	0.6937	0.8948
GB	0.9092	0.9376	0.6350	0.8369
LXGB	0.8901	0.9249	0.7484	0.8576
Stacked-ERT	0.8901	0.9283	0.7524	0.8587
Stacked-GB	0.8947	0.9287	0.7630	0.8649
Stacked-LXGB	0.9180	0.9464	0.6825	0.8588

Table 4.4. Supervised model results

The two best models are both extreme randomized trees: the first uses balanced subsampling -ERT-ss- (i.e., for every random sample used during the iteration of the trees, the sample is balanced by using weights inversely proportional to class frequencies), and serves here as our base model; the second uses an ADASYN oversampling method-ERT-os- and serves as our control model.

4.4.2 Investigation Office Validation

At the outset, we randomly set aside 10% of the data (30,317 claims). In this final step, we want to go further and examine these initial claims as test data. Our results are shown in Table 4.5.

As can be appreciated, the control model (Table 4.5b) has a recall of 97% while the base model (Table 4.5a) has an impressive recall of 100%. However, the real

Original Value	Prediction	Cases
NON-INVESTIGATED	NON-FRAUD	29.631
FRAUD	NON-FRAUD	0
NON-INVESTIGATED	FRAUD	415
FRAUD	FRAUD	271

(a) ERT-ss Robustness Check.

Original Value	Prediction	Cases
NON-INVESTIGATED	NON-FRAUD	29.656
FRAUD	NON-FRAUD	8
NON-INVESTIGATED	FRAUD	390
FRAUD	FRAUD	263

(b) ERT-os Robustness Check.

Table 4.5. Model Robustness Check.

added value depends on the non-investigated fraud cases, i.e., cases not previously detected but which would boost our results if shown to be fraudulent (non-investigated predicted as fraud). We, therefore, sent these cases to the IO for analysis.

The IO investigated 367 cases (at the intersection between the ERT-ss and ERT-os models). Two fraud investigators analyzed each of these cases, none of which they had previously seen as the rule model had not detected them.

Of these 367 cases, 333 were found to present a very high probability of being fraudulent. This means that only 34 could be ruled out as not being fraudulent. Recall that from the original sample of 415 cases, the fact that 333 presented indications of fraud means we have a precision of 88%. In short, we managed to increase the efficiency of fraud detection by 122.8%. These final outcomes are summarized in Table 4.6.

Original Value	Prediction	Cases
NON-INVESTIGATED	NON-FRAUD	29.631
FRAUD	NON-FRAUD	0
NON-FRAUD	FRAUD	(415 - 333) = 82
FRAUD	FRAUD	(271 + 333) = 604

Table 4.6. Base Model Final Results

4.4.3 Dynamic Learning

One of the challenges in fraud detection is that it is a dynamic process which can change its patterns over time. A year later, we retest the model with new data. We now have 519,921 claims to evaluate. We initially start out with a similar proportion of fraud cases (0.88%)- we are now able to train with 4,623 fraud cases to further improve results.

First, we recalculate the unsupervised algorithm, getting a Cluster-Score of 96.89%. As can be seen in Table 4.7, Cluster 0 contains almost every normal case. On

the other hand, we can clearly distinguish two fraud clusters: Cluster 1, in which 99.31% are fraud cases, and Cluster 2, in which 97.36% are fraud cases. Our 50% threshold therefore becomes insignificant again.

Clusters	Fraud	Percentage
0	0	99.4%
0	1	0.6%
1	0	0.7%
1	1	99.3%
2	0	2.6%
2	1	97.4%

Table 4.7. Oversampled Unsupervised Mini-Batch K-Means

Using the Extreme Randomized Subsampled approach (ERT-ss) and the Extreme Randomized oversampled with ADASYN (ERT-os), and the Stratified 5-fold cross validation approach we retrain the model. Table 4.8 shows the main results.

PERIOD	Jan15-Jan17	Jan15-Jan18
Claims	303,166	519,921
Observed Fraud	2,641	4,623
Cluster Score	96.59%	96.89%
Recall Score ERT-ss	97.34%	96.31%
Precision Score ERT-ss	67.18%	89.35%
F-Score ERT-ss	89.32%	94.84%
Recall Score ERT-os	96.47%	96.44%
Precision Score ERT-os	69.37%	92.18%
F-Score ERT-os	89.48%	95.56%

Table 4.8. Base Model with the machine-learning process applied

The base model greatly improves the homogeneity of the fraud and non-fraud clusters. In particular, it provides a gain of 33% in the precision score and of 6.2-6.8% in the F-Score.

4.5 Conclusion

This chapter has sought to offer a solution to the problems that arise when working with highly unbalanced data-sets for which the labeling of the majority of cases is unknown. In such cases, we may dispose of a few small samples that contain highly valuable information. Here, we have presented a fraud detection case, drawing on

the data provided by a leading insurance company, and have tested a new methodology based on semi-supervised fundamentals to predict fraudulent property claims.

At the outset, the Investigation Office (IO) did not investigate many cases (around 7,000 cases from a total of 303,166). Of these, only 2,641 were actually true positives (0.8% of total claims), with a success rate of 48%. Thanks to the methodology devised herein, which continuously readapts to dynamic and changing patterns, we can now investigate the whole spectrum of cases automatically, obtaining a total recall of 96% and a precision of 89-92%. In spite of the complexity of the initial problem, where the challenge was to detect fraud dynamically without knowing anything about 99.2% of the sample, the methodology described has been shown to be capable of solving the problem with great success.

4.6 Appendix. Practical Example

Imagine we have the following output from an unsupervised model:

Class	Label
0	1
0	2
0	3
0	1
1	2
1	2
1	3
0	3
0	3
0	2
0	1
0	3
1	2
0	1
1	2

Table 4.9. Class and Labels

The classes represent fraud (=1) and unlabeled (=0). The output label is the clustering label. As can be seen, just 33% of cases represent detected fraud. If we group the class by clusters:

Label	Class	Subtotal Class	Total
1	0	4	4
1	1	0	4
2	0	2	6
2	1	4	6
3	0	4	5
3	1	1	5

Table 4.10. Grouping Labels and Classes

As is evident, the fraud class tends to be assigned to the second cluster.
First we calculate C1.

$$C1 = \frac{\frac{0}{4} * 0 + \frac{4}{6} * 4 + \frac{1}{5} * 1}{5} = 0.5733$$

Then we calculate $C2$ using a similar formula.

$$C2 = \frac{\frac{4}{4} * 4 + \frac{2}{6} * 2 + \frac{4}{5} * 4}{10} = 0.7867$$

As can be seen, C1 gives worst results as its core group (group 2) is quite contaminated (66% of observations actually correspond to cases of fraud). This effect represents 93% of the total effect. The effect of mismatching the core group (1/5) is negligible, which stresses the importance of constructing a strong core group.

This conclusion is notorious in the case of C2. Non-identified classes are highly robust in two groups (1 and 3).

If we calculate the CS with $\alpha = 2$ (balanced C1 and C2) we obtain:

$$CS = 0.6062$$

which is a value very close to 0.5733. This formula allows us to balance our results, giving greater weight to the lower score. We should stress we want both good and balanced scores; thus, C1=0, C2=1 is not the same as C1=0.5, C2=0.5. Indeed, the former returns a CS=0. We compare the mean with the CS in Table 4.11.

$C1$	$C2$	$Mean$	CS
0.0	1.0	0.5	0.00
0.1	0.9	0.5	0.18
0.2	0.8	0.5	0.32
0.3	0.7	0.5	0.42
0.4	0.6	0.5	0.48
0.5	0.5	0.5	0.50
0.6	0.4	0.5	0.48
0.7	0.3	0.5	0.42
0.8	0.2	0.5	0.32
0.9	0.1	0.5	0.18
1.0	0.0	0.5	0.00

Table 4.11. $C1$ and $C2$ Combinations

As can be seen, we obtain the same unbalanced scores as the balanced outcomes for the mean score. CS penalizes the unbalanced scores. This is why we obtain different results with the same proportions.

However, we can make adjustments in terms of the relevance we attach to each group. If we raise α, we penalize the C2 results, and vice versa.

What happens if we choose $\alpha \geq 2$?

$C1$	$C2$	$\alpha = 2$	$\alpha = 4$	$\alpha = 6$
0.1	0.9	0.12	0.11	0.10
0.2	0.8	0.24	0.21	0.20
0.3	0.7	0.34	0.31	0.30
0.4	0.6	0.43	0.41	0.40
0.5	0.5	0.50	0.50	0.50
0.6	0.4	0.55	0.58	0.59
0.7	0.3	0.55	0.65	0.68
0.8	0.2	0.50	0.68	0.74
0.9	0.1	0.35	0.61	0.74

Table 4.12. $C1$ and $C2$ Combinations with $\alpha \geq 2$

As is evident, we obtain two effects. First, while C1 increases, CS also increases (although C2 decreases at the same rate). But the effect present in the balanced case now extends further. When we are at C1=0.7, the balanced effect tends to reverse

the situation or to slow the increasing rate. The second effect is that the score curve tends to a linear curve while we increase α. CS is now depending more strongly on C1 being higher; while the higher α the stronger C1.

5 Risk Categorization and Self-Reported Mechanisms in Automobile Insurance Markets

5.1 Introduction

In automobile insurance markets, companies face a severe problem of asymmetrical information during the underwriting process: they know next to nothing about their potential new customers, while the latter might tend to underreport prior claims when switching to a new company. In these markets, risk classification is generally explained by adverse selection which is a result of asymmetric information between insured and insurers. The insured are a heterogeneous group that has more information than the insurer, who is unable to differentiate between risk types. Indeed, it is a costly process for the company to detect who the high risk individuals are, and the latter have no incentives to reveal their true nature. This results in risk pooling (Arrow, 1963), which is necessarily inefficient as it averages the price between the low- and high-risk insured (Akerlof, 1970). There is a chance that the costs and claims will be higher than the premium paid by the customer. The opposite is to overrate the premium, and thereby becoming noncompetitive in the market and reducing the firm's amount of customers.

In one period contracts, basic insurance theory suggests that risky customers will not reveal their true nature and, therefore, a suboptimal Pareto equilibrium with an average premium will be reached if no additional incentives are imposed (see Rothschild and Stiglitz, 1976; Stiglitz, 1977). In this chapter we seek to address the following questions: Are all "bad risks" pretending to be "good risks" as insurance theory suggests? Or is it more nuanced, in that only a subset misreport their history? How relevant is "misreporting" in predicting risk, i.e. are "misreporters" particularly risky? And, is there any insight into what type of consumers are likely to misreport? Using self-reported data and observable data on potential customers from a leading insurance company in Spain, we find that combining self-reported data with observable characteristics allow us to reach an equilibrium that is close to

the public information equilibrium.

The first and most common mechanism to reduce asymmetry information is the **self-selection mechanism:** Insureds with different risk types self-select themselves among a menu of policies offered by the insurers. The result is a Pareto improvement compared to the single contract solution with an average premium (see Crocker and Snow, 2000). Based on the hypothesis that the high-risk individuals opt for higher coverage, empirical literature have evaluated conditional correlations between insurance coverage and subscribers' ex-post risk. The issue with this correlation, however, is that it fails to address the problem of unobserved data, as revealed ex-post the subscription process. Even though we assume that the coverage is a proxy of the policyholders' risk, there are additional unobserved differences among the insured (e.g., risk aversion, precaution levels, etc.). Furthermore, the fact that there is a correlation between coverage and risk does not mean that adverse selection is assured as it might also reflect moral hazard (which could nullify the effect). Finally, this correlation is not always conclusive (Dahlby, 1983, 1992; Chiappori and Salanié, 2000; Saito, 2006).

A second well-known mechanism is **risk categorization** which is based on variables that are costless to observe and are correlated with the unobservable risk of loss. For example, age, gender, type of car, etc. In this case, Pareto improvements are obtained by using imperfect signals to categorize risks (Hoy, 1982; Crocker and Snow, 1986). A well-known example is Dionne and Vanasse (1992) who find that young male drivers are riskier than young female drivers, using data from Quebec automobile drivers. However, this mechanism may generate an unfair price to the insured driver as it is based on imperfect signals. Several authors have mentioned this issue of underwriting and uncertainty. As early as 1982, de Wit pointed out that the process of setting the premium in non-life insurance was much harder due to the fact that using external data or shared company data is not always possible. The insured premium is therefore calculated only by available and measurable factors. In the end, what usually happens is that companies rely on simple judgment or decision rules. Even more, as mentioned by Kunreuther et al. (1995), the pricing risk tends to be higher in the existence of ambiguity about the probability of events. It is very common that private information, hidden within the applicant pool (high-risk observables as age, license years, etc.), explains a great part of the rejections.

However, if we assume that not all risky potential customers misreport as insurance theory suggests, we could combine this information with the traditional risk categorization method to solve its inefficiencies when predicting risky potential customers. Our main objective in this chapter is, therefore, to theoretically and empirically evaluate the combination of risk categorization tools and self-reported data in predicting risky customers ex-ante the policy is signed. Ideally, companies would

want to know new customers' unobserved characteristics (such as their past performance in other companies, reliability, etc.) to be able to detect high-risk customers and to distinguish them from those with good records. As we are usually unable to know the past behavior of a new customer, we propose mixing two mechanisms: First, observable data from the potential new customer and historical data from the internal customers as a proxy of the unknown characteristics; second, self-reported data about past performance behavior.

The Spanish insurance market offers us a good opportunity to evaluate the questions mentioned earlier and the ability to predict risky customers. Leading companies subscribe to a private service where they can share historical data about customers. Thus, when a new customer switches to another company, prior claims can be checked. The data comprises responses to only a few questions, but they are very precise. Specifically, they provide information about the number of years as insured and the number of previous claims. We can then use this information as validation of our methodology. Additionally, in the underwriting process, the policies that are offered are based, in part, on the responses to a number of questions by potential new customers. These questions also refer to years as insured and previous performance. If people were always to tell the truth about their own risk (i.e. revealing unobserved characteristics), it would be easy to set a fair price. As we observe below, only a very small number of people misreport (5.5% in our data-set), while those that lie increase the premiums offered to the truthful and, in some cases, this might even deter them from contracting with said company.

In doing so, we propose a new methodology for predicting risk. As companies have little information about new customers (but a considerable amount about their own), and only a small percentage lie about their true risk status, we show below that we can reduce the asymmetric information problem to one of anomaly detection. To do so, we propose applying deep variational autoencoder (VAE) models, that is, representations of neural networks, based on the idea of compressing and decompressing data. VAE models learn from a compressed representation of the data and from the latent variables obtained from the input. We expect the VAE model to minimize the asymmetric information problem by using a database containing only observables about new customers and internal customer data. In this way we seek to construct a model that can detect anomalies during the underwriting process and, thus, allow a fair price to be set as underreported claims should have been detected.

We conclude that combining self-reported data with risk categorization leads to two main outcomes: First, we are capable of predicting riskier customers ex-ante the policy is subscribed, something which we are not capable of when using only observable data. Second, we show that the most influential variables accounting for adverse selection are self-reported years as insured, cluster constructed variables

related to customers' zip code and characteristics, if the insured was the owner and first driver in the policy, if the insured's age was higher than 65, if the customer was male or female and the number of license years. The evidence we present is supported by empirical findings, which reveal that a great proportion of high risk individuals do not misreport their risk nature. Additionally, we present a theoretical model which helps explaining how combining self-reported data with risk categorization improves the potential profit for the insurer monopoly.

5.2 Literature Review

There is a vast literature on asymmetric information in insurance markets that has been strongly influenced by contract theory (Rothschild and Stiglitz, 1976; Stiglitz, 1977; Wilson, 1977; Miyazaki, 1977; Spence, 1978; Hoy, 1982; Kunreuther and Pauly, 1985; Dionne and Lasserre, 1985, 1987; Hellwig, 1986, Cooper and Hayes, 1987; Hosios and Peters, 1989; Nilssen, 2000; Dionne, 1992; Dionne and Doherty, 1994; Fombaron, 1997, 2000; Crocker and Snow, 1985, 1986, 2013). Several mechanisms have been proposed to reduce the inefficiency associated with adverse selection. The first and most common is the self-selection mechanism, where insureds with different risk types self-select themselves among a menu of policies offered by the insurers. The result is a Pareto improvement compared to the single contract solution with an average premium.

With the increasing availability of data (especially, in insurance companies), empirical evidence of adverse selection has acquired relevance. Several studies have evaluated conditional correlations between insurance coverage and subscribers' ex-post risk, based on the hypothesis that the high-risk insured opt for higher coverage. Thus, many authors report a positive correlation between risk and coverage (Brugiavini, 1993; Puelz and Snow, 1994; Chiappori, 1994; Dionne et al., 1999; Philipson and Cawley, 1999; Richaudeau, 1999; Dionne et al., 2000; Cardon and Hendel, 2001, Chippori et al., 2002; Finkelstein and Poterba, 2002, 2004; Davidoff and Welke, 2004; Cohen, 2005; Finkelstein and McGarry, 2006; Chiappori et al., 2006; He, 2009; Einav et al., 2010; Cohen and Siegelman, 2010).

A second type of mechanisms proposed to solve this resource allocation problem relies on multi-period contracts theory. Long-term contracting adjust ex-post insurance premiums or coverage to past behavior. Several studies have discussed the static notion of asymmetric information and introduce the notion of dynamic data to handle the issue of asymmetric learning (Abbring et al., 2003a, 2003b). Based on historical claims data, we can expect a positive correlation between past and fu-

ture claims, which reveals the policyholder's risk type[1]. Both insured and insurer may garner information about a policyholder's risk type. Hendel and Lizzeri (2003) find strong evidence of dynamic learning based on long-term contracts with a commitment to renew. Finkelstein et al. (2005) and Finkelstein and McGarry (2006) provide support for this evidence by using data on long-term care insurance. Furthermore, the learning process may not be symmetrical, and the policyholder can gain an advantage over the insurer. For instance, Cohen (2005) reports a positive correlation between coverage and risk for new customers with three or more years of driving experience, but finds no-correlation amongst new customers with less experience.

However, dynamic learning can only solve the lack of information problem ex-post the contract is signed, i.e., companies have to wait until policy renewal to resolve the problem. Additionally, companies are reticent to share past information about their own customers. Given that new customers tend to underreport prior claims when switching to a new insurer, each insurance company does in fact have an informational advantage over other companies as regards repeat customers. However, as D'Arcy and Doherty (1990) point out, insurance companies do not allow intermediaries to sell private information about their customers, and so by retaining information hold on to their market power. This practice has also been studied by Cohen (2008) who finds evidence consistent with insurance companies retaining private information about repeat customers. She demonstrates that companies obtain higher profits from repeat customers who have good records, and that these profits rise the longer the customer remains with the company. During the course of a contract, a customer's risk level may be revealed and this enhanced ability to determine an appropriate premium risk should result in higher profits. In this way, companies can discriminate their prices so that, eventually, customers with good records stay and those with bad records switch to another company. The problem remains, however, insofar as the company to which a customer is switching does not have this information and cannot, therefore, distinguish between those with bad records and those that switch due to exogenous factors.

A third type of mechanisms are those related with **categorization of risks**. In this case, Pareto improvements are obtained by using imperfect signals to categorize risks. Hoy (1982) is the first to theoretically analyze the effect of risk categorization in the market equilibrium, however, his conclusions about the usefulness of the method are ambiguous. Crocker and Snow (1986) compare utility-possibility frontiers with and without risk categorization, and they find that prior categorization results in a Pareto improvement below a specific threshold. Empirically, private

[1] Other relevant studies are Israel (2004) and Dionne et al. (2013).

information and rejections in insurance was studied by several authors. High-risk observable characteristics are usually applied by insurers to deter potential high-risk insureds. Murtaugh et al. (1995), in long-term care insurance, estimate that between 12-23% of potential customers would be rejected if everyone applied at the age of 65 (20-31% if applied at 75). Hendren (2013) reports that between 2007 and 2009, 1 in 7 applicants to the four largest health insurance companies in the United States were rejected, and shows how the existence of private risk information implied that insurance companies rejected people only based on observables (for three different health insurance markets). There is a vast literature that relates risk classification and observable characteristics. Puelz and Kemmsies (1993), Lemaire (1995) and Doerpinhaus et al. (2008) evaluate the impact of gender and other demographic variables on premium pricing. Age and risk classification is another well-studied relation (Brown et al., 2007; Braver and Trempel, 2004; Tefft, 2008). Dionne and Vanasse (1992) find strong evidence between risk and young males, and also between risk and classes of driver's license.

Several authors propose different classification systems to classify risk in insurance: statistical measures (Tryfos, 1980; Freifelder, 1985; Lemaire, 1985; Samson, 1986; Dionne and Vanasse, 1992), linear models (Samson and Thomas, 1987; Ohlsson and Johansson, 2010; Bortoluzzo et al., 2011; David, 2015), clustering techniques (Williams and Huang, 1997; Smith et al., 2000; Yeo et al., 2001). However, these kinds of models have several limitations regarding: solving non-linear relations, the existence of too many variables, and high dispersion (which is quite common for insurance databases). A few recent studies have started to apply machine-learning and data mining techniques to claims and risk (Gepp et al., 2012; Guelman, 2012; Liu et al., 2014; Yang et al., 2015; Kašćelan et al., 2015). In our case, we propose using a deep Variational Autoencoder approach. Autoencoder networks have traditionally been used in image classification (Tan et al., 2010; Krizhevsky and Hinton, 2011; Hinton et al., 2011; Walker et al., 2016; Pu et al., 2016; Theis et al., 2017); however, they have recently been used more frequently with structure data, especially in relation to anomaly detection problems (Dau and Song, 2014; An and Cho, 2015; Andrews et al., 2016; Zhai et al., 2016; L. Paula et al., 2016; Zhou and Paffenroth, 2017; Cozzolino and Verdoliva, 2017; D'Avino et al., 2017; Schreyer et al., 2017). The main advantages of using variational autoencoder networks are: Firstly, they can reduce data to their true nature, cleansing them of any undesired features and noise. Here, we have a large data-set containing many variables about internal customers, but as we do not know which are relevant for predicting risky potential customers, VAE should be able to help us address this problem. It works by reducing the number of nodes through hidden layers, so that we can extract the actual features representing the data. Secondly, existing outlier detection approaches

are usually based on notions of distance and density, meaning they work in the original data's space. As a result, they tend to underperform when applied to nonlinear structures. To solve this issue, a deep autoencoder model can be transformed into a powerful semi-supervised outlier detection model.

5.3 Theoretical Model

The model developed by Rothschild and Stiglitz (1976) and further elaborated by Stiglitz (1977) introduces a risk neutral private monopoly that offers insurance coverage (β_i) for an insurance premium (α_i) in a single-period contract and under public information. For simplicity, there are three types of risk ($i \in \{L, M, H\}$). Initially, we assume that risk type H and M have the exact same level of risk. Thus, $p_H = p_M > p_L$. Each insured owns a risky asset with monetary value $D(x)$ which depends on two possible states of the world $x \in \{n, a\}$: a represents the accident state with probability p_i, and n represents the no accident state with probability $1 - p_i$, such that $D(a) = 0$ in state a and $D(n) = D$ in state n.

The expected utility of the insured under the contract $C_i = \{\alpha_i, \beta_i\}$ and initial wealth W_0, is equal to:

$$V(C_i|p_i) = p_i U(W_0 - D - \beta_i) + (1 - p_i)U(W_0 - \alpha_i)$$

which is strictly concave and satisfies the von Neumann-Morgenstern axioms[2]. Therefore, with public information and without transaction cost, the problem can be summarized as:

$$\max_{\alpha_i, \beta_i, \lambda_i} \sum q_i[(1 - p_i)\alpha_i - p_i\beta_i]$$

Subject to the participating constraint $V(C_i|p_i) \geq V_i^0$ (the monopoly can extract all the consumer surplus). In presence of public information about insureds' risk, it can be proved that insureds get full insurance coverage ($\beta_i^* = D - \alpha_i^*$), that there is no consumer surplus, and that $\alpha_L < \alpha_M = \alpha_H$. We illustrate this solution in Figure 5.1. As the monopoly can distinguish between insureds, it offers one of three contracts: $\{C_L^*, C_M^*, C_H^*\}$ to which insureds are indifferent between the offered contract and the self-insurance contract C_0.

Hereinafter, we assume the existence of private information about insureds' risk types. Theory suggests that if the monopoly offers the same contract to different risk types, it is rational for both, H and M, to move to a contract C_L and, the consequence is a pooled equilibrium where the monopoly cannot distinguish among

[2] $U' > 0, U'' < 0$.

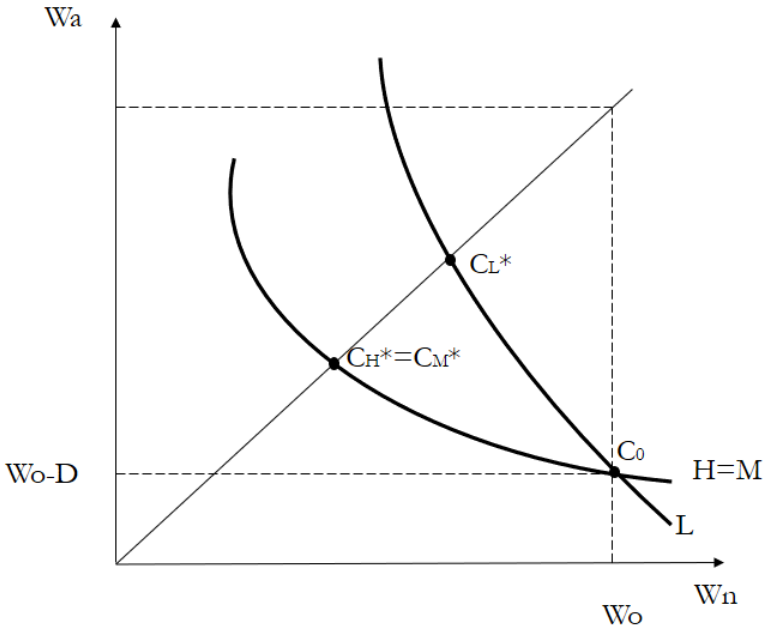


Figure 5.1. Public Information Equilibrium

individuals. In this case, introducing incentives will allow low risk individuals to reveal their true nature by subscribing a policy with limited coverage at a lower price. Formally, a self-selection constraint is defined as:

$$V(C_i|p_i) \geq V(C_j|p_i) \forall i,j \in \{L,M,H\}, \quad i \neq j$$

Basically, it guarantees that insured i prefers his or her own contract over any other contract. From Figure 5.1, it is clear that if the contracts $\{C_L^*, C_M^*, C_H^*\}$ are offered, both L and H will pool in C_L^*. Therefore, in order to avoid a pooled equilibrium, the monopoly should offer a combination of contracts that does not encourage high risk individuals to pool with low risk individuals. For instance, the insurer can offer a set of contracts where high risk insureds are indifferent between the low risk contract and their own contract. And, if they are indifferent, low risk insureds will strictly prefer their own contract. Therefore, the maximization problem is defined as:

$$\begin{aligned} \max_{\alpha_i,\beta_i,\lambda_i} \quad & \sum q_i[(1-p_i)\alpha_i - p_i\beta_i] \\ \text{s.a.} \quad & V(C_j|p_j) \geq V(C_L|p_j) \quad j = H,M \\ & V(C_L|p_L) \geq U_L^0 \end{aligned} \tag{5.1}$$

Figure 5.2 shows the solution of this maximization. The monopoly offers $C_H^{**} = C_M^{**}$ to H and M who are indifferent between this contract and the contract C_L^{**}

offered to L. It also shows that $C_H^{**} = C_M^{**}$ is strictly preferred to $C_H^{*} = C_M^{*}$, which means that the participation constraint is not binding for these groups (and, therefore, their consumer surplus is not fully extracted). Additionally, insureds L get no surplus but they strictly prefer their contract to the one offered to M and H individuals. Finally, individuals L get a lower coverage and a lower price compared to the contract C_L^{*}.

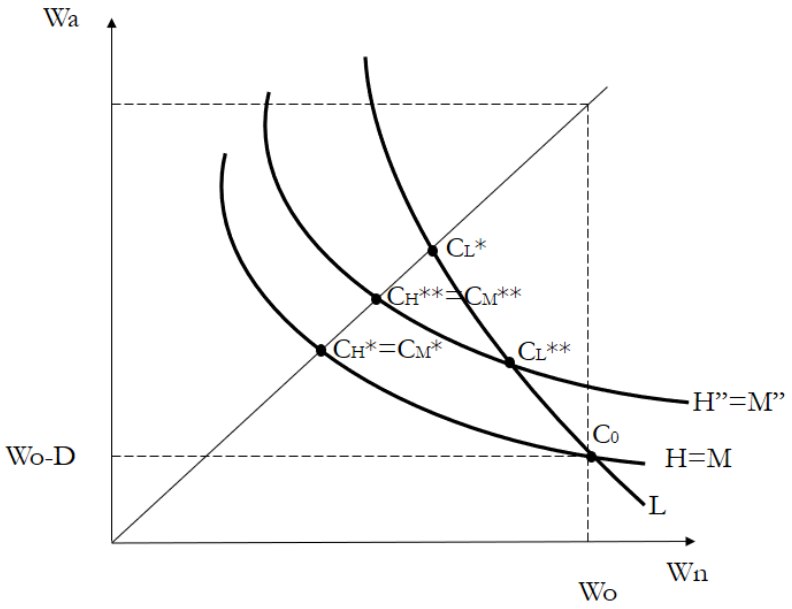


Figure 5.2. Private Information Equilibrium

In conclusion, monopoly profits are increased compared to a situation without any revelation mechanism, which does not necessarily correspond to the best risk allocation or the maximum profit allocation, as good risk individuals are partially covered (they may even refrain from purchasing any insurance at all). Specifically, it can be demonstrated that $\beta_{H,L} = D - \alpha_{H,L}^{**}$ and $\beta_L < D - \alpha_L^{**}$.

5.3.1 Self Selection Mechanism

This subsection presents a extension of the model first developed by Rothschild and Stiglitz (1976) and extended by Stiglitz (1976). As we will test empirically below, in certain cases one might find that some high risk individuals reveal their true nature (we will discuss later why those insureds might be willing to self-report their true nature). For now, we know that insurance theory suggests that with private information and without incentive constraints, high risk individuals choose to pool with low risk individuals because by doing this they can get full coverage at a lower price. Then, suppose that we have the same groups of insureds but one group (the M

insureds), for some reason (idiosyncratic reasons, fear to future penalization, differences in risk aversion, etc.), reveals its true nature and, therefore, the monopoly has public information about it. In this case, the monopoly can increase its benefits by offering a new contract to which this group is indifferent between the new contract and self-insurance (see Figure 5.3).

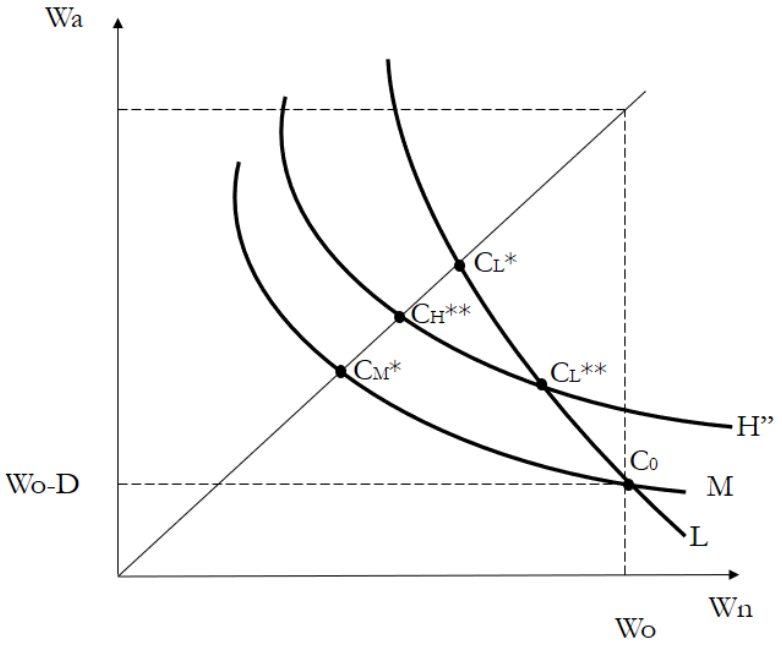


Figure 5.3. Semi-Private Information Equilibrium

In this situation, the monopoly must still separate between H and L and, thus, the maximization profit problem can be defined as:

$$\begin{aligned} \max_{\alpha_i,\beta_i,\lambda_i} \quad & \sum q_i[(1-p_i)\alpha_i - p_i\beta_i] \\ \text{s.a.} \quad & V(C_j|p_j) \geq V(C_L|p_j) \quad j = H, M \\ & V(C_L|p_L) \geq U_L^0 \end{aligned} \tag{5.2}$$

Proposition 1. *In the presence of semi-private information, i.e., a situation where the monopoly knows the nature of some high-risk individuals (e.g., insureds M), the optimal one-period contract equilibrium has the following characteristics:*

- *The monopoly offers C_M* such that individuals M have no consumer surplus: $V(C_M^*|p_M) = V_M^0$.*
- *The monopoly offers C_L^{**} to insureds L and it also extracts the consumer surplus: $V(C_L^{**}|p_L) = V_L^0$.*

- *Insureds H strictly prefer C_H^{**} to C_M* and they are indifferent between C_L^{**} and C_H^{**}. Consumer surplus is positive: $V(C_H^{**}|p_H) > V_0^H$.*

- *Both, H and M get full coverage equal to $\beta_H^{**} = D - \alpha_H^{**}$ and $\beta_M^{*} = D - \alpha_M^{*}$. Individuals L get partial coverage (or no coverage) at a lower price, more precisely, $\beta_L^{**} < \beta_H^{**} < \beta_M^{*}$ and $\alpha_L^{**} < \alpha_H^{**} < \alpha_M^{*}$.*

- *The monopoly profit is in this case higher than in the full private information equilibrium but lower than in the public information equilibrium: $\pi(C_L^{**}, C_M^{**}, C_H^{**}) < \pi(C_L^{**}, C_M^{*}, C_H^{**}) < \pi(C_L^{*}, C_M^{*}, C_H^{*})$.*

In conclusion, the monopoly increases the profits compared to the private information case. However, partial coverage (and even no coverage) and offering a lower price to low risk individuals is still necessary to separate between high and low risk individuals.

5.3.2 Risk Categorization

The monopoly can combine information from some of the high risk individuals that have revealed their true nature and a risk categorization mechanism to separate high risk individuals from low risk individuals. The risk categorization theory suggests that imperfect information such as observable characteristics can be used to obtain Pareto improvements for resource allocation (Hoy, 1982; Crocker and Snow, 1986).

Suppose for example that we have derived two possible groups from the observable characteristics of self-selected insureds. Then, we assume that insureds differ in characteristics that are costless to observe and which are also correlated with the unobservable risk. By using this information, we separate between: Group A with a proportion of high risk individuals w_H^A and low risk individuals w_L^A, and group B with a proportion of high risk individuals w_H^B and low risk individuals w_L^B. The observable characteristics determine the proportion of high risk individuals is higher in the group A than in the group B, that is, $w_H^B > w_H^A$ and $w_L^A > w_L^B$ ($w_H^B = 1 - w_L^B$ and $w_H^A = 1 - w_L^A$). Bearing this in mind, we derive the proposition below. For simplicity, we add a group C in which we know all individuals are type M individuals. The size of each group is defined as: q_L, q_M and q_H, respectively.

Proposition 2. *In the presence of semi-private information, it is possible to reach the Pareto optimal equilibrium under perfect information if there are characteristics that are costless to observe and which correlate perfectly with the unobservable risk of loss.*

To demonstrate this, we first define the profit definition under perfect information:

$$pi^* = q_L\pi(C_L^*|p_L) + q_M\pi(C_M^*|p_M) + q_H\pi C_H^*|p_H)$$

If the monopoly knows that the characteristics are highly correlated with the risk, then it can offer group A (with a high proportion of low risk individuals) the contract offered to L in the perfect information case (C_L^*), and it can offer group B contract C_H^*. Thus, the profits with risk categorization and self-reported risk can be formally defined as:

$$\begin{aligned}\pi^{RC} = q_A[w_H^A\pi(C_L^*|p_H) + (1-w_H^A)\pi(C_L^*|p_L)] + \\ q_B[w_H^B\pi(C_H^*|p_h) + (1-w_H^B)\pi(C_H^*|p_L)] + q_C\pi(C_M^*|p_M)\end{aligned} \tag{5.3}$$

If the categorization is accurate: $w_H^A \to 0$, $w_H^B \to 1$, $q_A \to q_L$ and $q_B \to q_H$. Then, we can conclude that:

$$\pi^{RC} \to \pi^*$$

Proposition 3. *In the presence of semi-private information, there exist equilibria in which the monopoly can get higher profits than the private information equilibrium (but strictly lower than the perfect information equilibrium), if there are observable characteristics that are costless to observe and which do not correlate perfectly with the unobservable risk of loss. The stability of these equilibria finally depend on to which extent the risk categorization tool correlates with the risk.*

If the monopoly offers group A contract C_L^*, group B contract C_H^* and group C contract C_M^*, the monopoly profit can be defined as:

$$\begin{aligned}\pi^{CR} = q_A(1-w_H^A)\pi(C_L^*|p_L) + q_B(1-w_H^B)\pi(C_H^*|p_L) + \\ q_A w_H^A\pi(C_L^*|p_H) + q_B w_H^B\pi(C_H^*|p_H) + q_C\pi(C_M^*|p_M)\end{aligned} \tag{5.4}$$

However, if the signal is imperfectly informative, then, some L individuals are offered C_H^*. Thus, the participation constraint is violated because at this particular utility level, L does not participate. As a result, $\pi(C_H^*|p_L) = 0$, which implies that without a perfect signal, profits are strictly lower than the profits under perfect information.

We now focus on the next part of the proposition. That is, if there exist equilibria where imperfect information allows the monopoly to get higher benefits than with private information. That is:

$$\pi^{CR} \geq \pi^{**}$$

Which is equivalent to:

$$\begin{aligned}\pi^{RC} = \quad & q_A(1-w_H^A)[(1-p_L)\alpha_L^* - p_L\beta_L^*] \\ & + q_A w_H^A[(1-p_L)\alpha_L^* - p_H\beta_L^*] \\ & + q_B w_H^B[(1-p_H)\alpha_H^* - p_H\beta_H*] \\ +q_M[(1-p_M)\alpha_M^* - p_M\beta_M^*] \geq & \\ & \pi^* = q_L[(1-p_L)\alpha_L^{**} - p_L\beta_L^{**}] \\ & + q_M[(1-p_M)\alpha_M^* - p_M\beta_M^*] \\ & q_H[(1-p_H)\alpha_H^{**} - p_H\beta_H^{**}\end{aligned} \tag{5.5}$$

Rearranging and replacing $q_H = q_A w_H^A + q_B w_H^B$ and $q_L = q_A(1-w_H^A) + q_B(1-w_H^B)$:

$$\begin{aligned} q_A(1-w_H^A)[(1-p_L)(\alpha_L^* - \alpha_L^{**}) - p_L(\beta_L^* - \beta_L^{**})] \\ -q_B(1-w_H^B)[(1-p_L)\alpha_L^{**} - p_L\beta_L**] \\ +q_A w_H^A[(1-p_H)(\alpha_L^* - \alpha_L^{**}) - p_H(\beta_L^* - \beta_H^{**})] \\ +q_B w_H^B[(1-p_H)(\alpha_H^* - \alpha_H^{**}) - p_H(\beta_H^* - \beta_H^{**})] \end{aligned} \tag{5.6}$$

The first term is the profit from moving the L individuals in A from the contract offered in the private information case (C_L^{**}) to the perfect information contract (C_L^*). Since we know that $\beta_L^* = D - \alpha_L^*$ and $\beta_L^{**} < D - \alpha_L^{**}$, then $\alpha_L^* - \alpha_L^{**} > \beta_L^* - \beta_L^{**}$. In consequence, we can conclude (by logically assuming that $1 - p_L > p_L$) that this term is strictly positive.

The second term is the opportunity cost of the L types who previously were placed in group B and who now decided to not participate (which is strictly negative).

The third term represents the loss obtained by offering H types in A the C_L* contract instead of offering the C_H^* contract. As we previously stated, $\alpha^* < \alpha^{**}$ and $\beta_L^* > \beta_H^{**}$. Therefore, this term is also negative.

Finally, the last term is the profit of moving H individuals from C_H^{**} to the contract C_H^*. As $\alpha_H^* > \alpha_H^{**}$ and $\beta_H^* < \beta_H^{**}$, this term is positive.

For homothetic preferences we can derive that $\beta_L^* - \beta_H^{**} = \alpha_H^{**} - \alpha_L^*$ and $\beta_H^* - \beta_H^{**} = \alpha_H^{**} - \alpha_H^*$, and assuming that $q_A = q_B$, we can formally define:

$$\begin{aligned}(1-w_H^A)[(1-p_L)(\alpha_L^* - \alpha_L^{**}) - p_L(\beta_L^* - \beta_L^{**})] + w_H^B(\alpha_H^* - \alpha_H^{**}) \\ \geq (1-w_H^B)[(1-p_L)\alpha_L^{**\smile} p_L\beta_L^{**}] + w_H^A(\alpha_H^{**} - \alpha_L^*)\end{aligned} \tag{5.7}$$

Based on the above, we know that the first two terms are positive. This inequality

will always be satisfied if $w_H^B \to 1$ and $w_H^A \to 0$. That is, the equilibrium is sustainable if the group membership is sufficiently informative. In conclusion, we can reach the perfect information equilibrium with an imperfect signal if it is sufficiently informative.

5.4 Data

In this section we introduce the data that will be used to test, firstly, whether all high risk individuals pretend to be low risk individuals as basic insurance theory suggests; secondly, the advantage of using this information to predict ex-ante risk. Our data set not only contains self-reported data about past claims but also data provided by an external database which is used to corroborate what potential insureds have self-reported.

If our hypothesis that not all high risk individuals misreport is true, we expect that adding self-reported data to observable characteristics will strengthen our predictive power.

5.4.1 Description of the Data

The data comprises a sample of vehicle insurance policies that were offered to new customers in Spain by a leading insurance company during the first six months of 2018 (14,817 observations). Each observation represents an offer made to a new customer that may have been transformed into a policy or not.

Before the policy is subscribed, the potential customer provides information about his or her observable characteristics and about the vehicle that is to be insured. The company then asks a few questions about past performance, specifically about historical claims and number of years as insured. This process takes place before the customer is informed of the final price.

During the period of analysis, leading insurance companies in Spain also have access to a source of unobservable risk data, which was provided by an external service. This database held information about a potential new customer's prior claims with other companies. Every time a new customer wanted to subscribe a new contract, this database was consulted using his or her national identification number.

In our case study, the company asked a potential customer a series of questions before offering a price. The answers given by the customer were later checked against the information available in the database provided by the external service. Customers were unaware of this process and so it provides us with a good opportu-

nity to identify customers who fail to reveal their true risk status. By comparing the answers provided by the new customers to the information held on the database, we can verify just how capable we are of identifying individuals that underreport their prior claims.

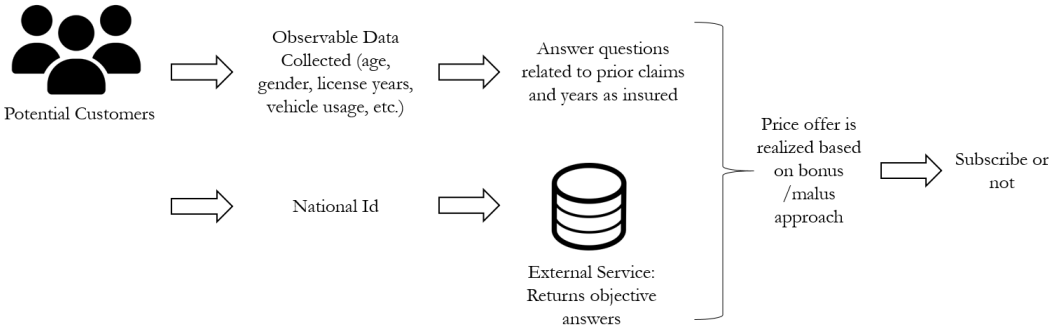


Figure 5.4. Policy Subscription

The company set the premium price as follows: After the potential customer had answered questions relating to number of years as insured and number of prior claims, a simulation bonus was calculated based on these answers. The external database was consulted and a corresponding bonus was generated. A final bonus/malus was then offered after contrasting the simulations in line with certain rules. Further details about customer responses and how the bonus was calculated are included below in subsection 4.2: Target Variable.

As discussed, one of our main objectives is try to predict, ex-ante, risky customers, especially, those individuals that misreport claims. As stressed, insurers usually do not know whether a new customer is hiding risk information or not, given that this information is essentially unobservable. We do however have access to an external source, but our aim herein is to find a way to avoid having to resort to using it. Therefore, in what follows, we only consult this database to control and evaluate our results.

Internal Data

As emphasized above, the main problem we face is that we have no information about unobservables, i.e. data related to the prior performance of new customers. However, we do have a great deal of rich, detailed information about internal customers. How might this help us predict risky potential customers? What we propose is using the information we have about active customers as a proxy for the new customers' past behavior. We can achieve this by segmenting the observables of the internal customers to detect which profiles are most likely to be risky. Thus, we use a database containing information about customers that signed a vehicle policy

between January 2013 and December 2017, and compare it with the characteristics of the company's potential customers during the first six months of 2018. In total we have 191,746 customers after sanitization, with more than 50 variables concerning the customer (gender, age, zip code, etc.), the vehicle (usage, type, brand, etc.), the intermediary (number of policies, number of claims, etc.) and past performance (number of policies, number of claims, number of guilty claims, claims cost, etc.). The goal is to create aggregated pools of observable data that can be matched with unobservables.

It should be borne in mind that each policy has a different life cycle, i.e., they do not all necessarily start and finish on the same date and are, therefore, not directly comparable. To rectify this, we normalize the past performance variables by the number of days during which the policy was active. For instance, in the case of two policies, each registering a single claim, but where one was signed a year later than the other policy, the older of the two is assumed to have a lower risk. Basically, we follow the formulation below:

$$x_i = \frac{x_i}{(N_{activepolicydays} + 1)/365}$$

Table 5.1 displays the variables in the internal data-set. It can be seen that the average customer is 48.93 years old and has been a customer for an average of 2.32 years, 67% of all customers are male and 80% are Spanish. In general, a customer has 1.51 policies (1.39 vehicle policies), and has been in possession of a driving license for 25.62 years. On average, a customer has 0.49 claims per year, of which 36% are guilty claims, with an estimated cost of 243.36 euros. 80% of all the policies sold by intermediaries are vehicle policies and, in general, for each, one claim is registered. Finally, the vehicles are worth an average of 21,854.20 euros, and are 9.62 years old.

Based on the observable variables in the internal customer database, we can pool types associated with different risk levels. To do so, we use unsupervised methods to cluster our observable data and then rank the clusters in terms of risk (past claims/cost). Thus, we expect to be able to capture riskier groups based on observables and to be able to match outcomes with the potential new customers' observables. Based on several evaluation metrics, we use a K-means++ (Arthur and Vassilvitskii, 2007) approach for obtaining four different categories of clusters: 1) ZIP code risk, 2) Intermediary risk, 3) Object risk and 4) Customer risk (see Appendix: Clustering Observable Risk Variables for a more detailed analysis).

Name	Group	Type	Description	mean	std
Number of claims	Claims	Integer	Number of claims per year	0.49	1.03
Number of vehicle claims	Claims	Integer	Number of vehicle per year	0.47	1.01
Claim cost	Claims	Float	Total claim cost per year	243.36	1,746.15
Refused claims	Claims	Integer	Number of refused claims	0.01	0.08
Guilty car claims	Claims	Integer	Total number of guilty car claims per year	0.37	0.84
Guilty claims percentage	Claims	Float	Percentage of guilty claims respect to total claims	36.73	45.64
Gender	Customer	Boolean	If insured gender is male equals 1, otherwise 0	0.67	0.47
Age	Customer	Integer	Customer age	48.93	13.27
Years as a costumer	Customer	Integer	The number of years as a costumer	2.32	1.46
Postal code	Customer	Integer	Postal code number	-	-
Id type	Customer	String	Customer national ID type	-	-
Birth date	Customer	Date	Birth date	-	-
Nationality	Customer	String	Customer nationality	-	-
Residence country	Customer	String	Customer residence country	-	-
Initial date as a costumer	Customer	Date	Initial date as a customer	-	-
Date of birth first driver	Customer	Date	Date of birth first driver	-	-
Date of birth second driver	Customer	Date	Date of birth second driver	-	-
License expedition date first driver	Customer	Date	License expedition date first driver	-	-
Quantity of policies	Customer	Integer	Total number of policies	1.51	1.13
Quantity of car policies	Customer	Integer	Total number of car policies	1.39	0.97
Policy initial date	Customer	Date	Policy initial date	-	-
Risk driver age	Customer	Boolean	If driver age is lower than 22 years old equals 1, otherwise 0	0.00	0.04
Risk second driver age	Customer	Boolean	If second driver age is lower than 22 years old equals 1, otherwise 0	0.01	0.09
License years	Customer	Integer	Number of years of the license	25.62	12.81
Risk license years	Customer	Boolean	If license years is lower or equal to 1, then 1, otherwise 0	0.00	0.06
Risk second driver license years	Customer	Boolean	If second driver license years is lower or equal to 1, then 1, otherwise 0	0.01	0.08
Nationality region	Customer	String	Nationality region (South America, West Europe, etc.)	-	-
Foreigner	Customer	Boolean	If customer is not Spanish equals 1, otherwise 0	0.20	0.40
Intermediary: number of vehicle policies	Intermediary	Float	Number of policies per year	500	793
Intermediary: number of vehicle claims	Intermediary	Float	Number of claims per year	400	586
Intermediary risk	Intermediary	Float	Number of claims over the number of policies	0.98	0.49
Vehicle intermediary risk	Intermediary	Float	Number of vehicle claims over the number of vehicle policies	1.02	0.50
Vehicle policy share	Intermediary	Float	Number of vehicle policies of the intermediary and the number of total policies	0.80	0.19
Vehicle usage	Object	String	Particular, rental, industrial, etc.	-	-
Vehicle value	Object	Float	Vehicle value in euros	21,854.20	11,455.52
Vehicle class	Object	String	Pick up, familiar, track, etc.	-	-
Vehicle aggregation	Object	String	Tourism, van, all terrain, etc.	-	-
Vehicle power	Object	Integer	Vehicle power	109.71	47.67
Vehicle brand	Object	String	Vehicle brand	-	-
Vehicle model	Object	String	Vehicle model	-	-
Vehicle category	Object	String	Particular, Motorcycle, Others	-	-
Vehicle fuel type	Object	String	Vehicle fuel type	-	-
Vehicle heavy	Object	Boolean	If vehicle weight is higher than 3,500 kg equals 1, otherwise 0	0.00	0.02
Vehicle age	Object	Integer	Vehicle age in years	9.62	7.27

Table 5.1. Internal Customer Data

Offer Data

These data include details of the policy terms offered to potential new customers. The information is very similar to that of our internal data and we can, therefore, use this database as a comparable source. The ultimate goal is to match the labels obtained above from the clusters with the observable variables associated with these offers. For instance, to a potential customer of certain characteristics (that is, age, gender, license year, etc.), we can assign him or her a risk level based on the cluster labels. The other information about the customer and his or her vehicle can then serve as our predictors.

5.4.2 Target Variable: The Definition of Risk

Once potential customers have reported details about their characteristics, the vehicle they wish to insure and the policy type they are interested in purchasing, the company applies its own specific rules in order to calculate the final price. Each policy has a technical price (or base price) and, as Chiappori and Salaniee (2000) explain, all insurers are required by law to apply a uniform price based on a "bonus/malus" approach. The premium can, therefore, be defined as the product price (technical price) plus a bonus coefficient:

$$Premium = 1 + bonus/malus$$

Each company has its own rules, but in general, the premium is closely correlated to past claims and, to a lesser extent, years as insured.

In the case of this particular company, it asked customers various questions before offering them a price. With the information obtained, it calculated a "simulated bonus" based on answers to the following questions:

1. Years as insured.
2. Years as insured in a previous company.
3. Guilty claims.
4. Guilty claims in the last three years.
5. Previous company.

However, before subscribing the policy, the company accessed an external database in which the leading insurers shared the same information regarding claims and years as insured. Using these data, the company also set an "adjusted bonus" for

the potential customer. For example, a person may had declared zero claims and so obtained a positive simulated bonus, but on consulting the database if the company found that, in fact, he had made five claims, then he received a negative adjusted bonus. A comparison of the two outcomes allowed the insurance company to decide whether to add a discount or a penalty to the technical premium. The company was then in a position to offer a policy based on the "emission bonus". The comparison considered two components – level and percentage – and follows the rules set out below:

1. If simulation bonus level $<$ adjusted level bonus, emission bonus $==$ simulation bonus.

2. If simulation bonus level $==$ adjusted level bonus, emission bonus $==$ simulation bonus $==$ adjusted bonus.

3. If simulation bonus level$-$ adjusted bonus level $== 1$ and adjusted bonus $\geq 30\%$, emission bonus $==$ simulation bonus.

4. If simulation bonus level $-$ adjusted bonus level $== 1$ and adjusted bonus $< 30\%$, emission bonus $==$ adjusted bonus.

5. If simulation bonus level $-$ adjusted bonus level > 1, emission bonus $==$ adjusted bonus.

If people always tell the truth, the price can be assumed to be fair. However, not every potential customer reveals his or her true nature, which is why insurance companies checked details to the external service. This external database held the true responses to the previously asked questions and, so, calculated the discount/penalty that had to be applied if the customer had not been truthful. In this way, the problem of risk categorization is reduced to simply detecting who has lied during the underwriting process.

Thus, in accordance with the insurance company, a person was not telling the truth when rules 4) and 5) did not hold, i.e., when the simulation bonus was replaced by a lower adjusted bonus. Our target variable is, therefore, 1, when 4) and 5) apply, which represents 5.5% of total offers. **This is the key point to understand our particular definition of risk: An algorithm detecting risk is actually an algorithm that can detect who has misreported, that is the potential customers who intentionally hide information about their true nature.**

Here, we focus only on simulation bonuses that are zero or positive, because the company's system does not allow us to continue the process with negative bonuses.

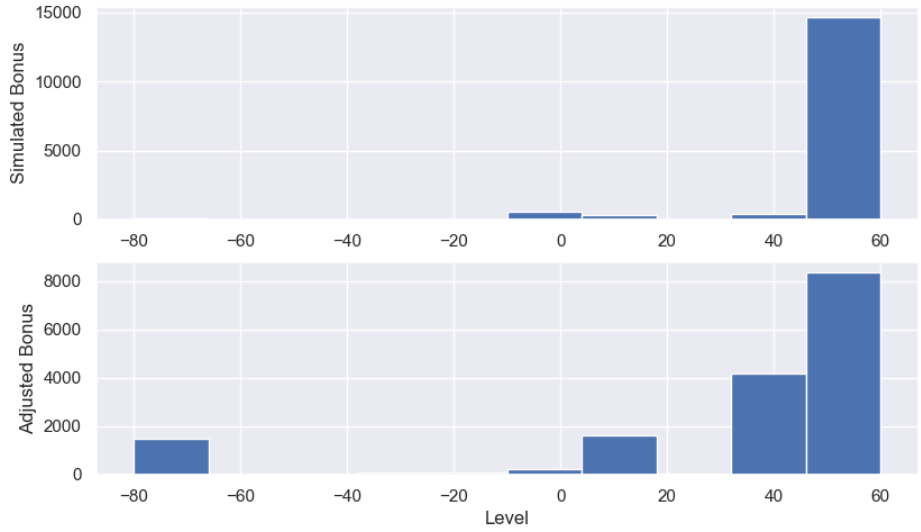


Figure 5.5. Simulated and Adjusted Bonus

In Figure 5.5, we show the simulated and adjusted bonus distributions. Simulated bonuses tend to be concentrated at 60%, and adjusted bonuses between 40-60%.

Based on the above rules, we obtain a target variable that indicates that 5.5% of the policy offers have had their simulation bonus modified (see Figure 5.6). This means that, in the case of almost 95% of the offers, checking the external database was unnecessary as the simulation bonus was already correct, i.e., potential customers had not lied. As such, the problem we face is detecting a small number of potential new risky customers, in other words, we have to address an outlier problem.

5.5 Methodology

We carry out our tests by taking advantage of a deep Variational Autoencoder model: A deep autoencoder is a representation of a neural network, trained by unsupervised learning. It permits a dimensional reduction to be applied in a hierarchical fashion and for learning to take place from reconstructions that are close to its original input. By successively stacking encoders, it is capable of obtaining more abstract features. The encoder and decoder are usually nonlinear mappings, which consist of several layers of neurons. The encoder maps the input vector (x) to a hidden representation which is then mapped back by the decoder to a reconstructed vector ($\hat{x}$). By reconstructing the data with low dimensions, we expect

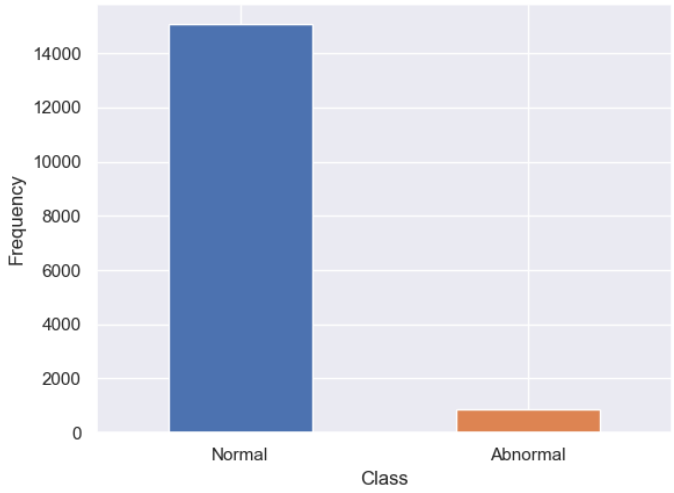


Figure 5.6. Anomaly Distribution

to obtain the data's true nature or the specific target we seek, especially as we are dealing with many variables without knowing which are relevant or not.

A major advantage here is that we can use the autoencoder as an outlier detection model. Reconstruction errors in the data points (the difference between the original value and the compress-decompress value) can be used as an anomaly score. The idea is to transform the model into a semi-supervised model, using only normal instances to train the autoencoder. Data points with high reconstruction errors are considered anomalies. After training, the autoencoder is able to reconstruct normal data whereas it is unable to do so when it encounters anomaly data, which it has not seen before.

We can also exploit here variational autoencoder (VAE) models, which are a modification of autoencoder models but with an added constraint on the encoded representation. More specifically, it is an autoencoder that learns a latent variable model from its input data. As such, the main difference is that the autoencoder is a deterministic model that does not use probability foundation, while VAE is a generative model. Instead of letting the neural network learn an arbitrary function, it learns the parameters of a probability distribution that models the input data. Figure 5.7 illustrates a schematic VAE model (a detailed analysis and validation of the model is presented in Appendix: Variational Autoencoder Model Validation).

We propose driving the VAE model as semi-supervised learning and comparing it with other machine-learning and deep-learning models. We split our data-set in the following way:

1. Training Set: 70% of the Normal data is used as the training set.

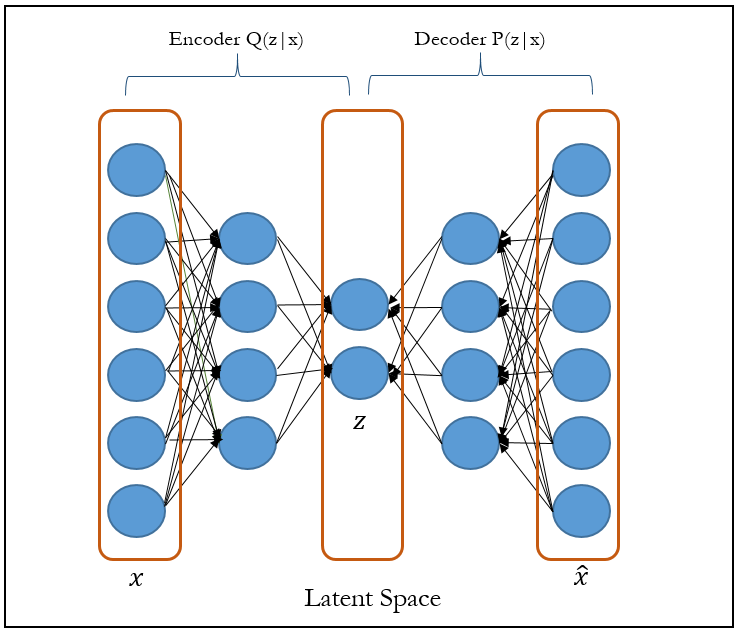


Figure 5.7. Deep Variational Autoencoder

2. Normal Test Set: Of the remaining 30%, we retain a percentage equivalent to the abnormal set as part of the test set.

3. Abnormal Test Set: The other part of the test set is made up of the abnormal cases. We therefore have a 50/50 test set.

4. Valid Test Set: The remaining normal data are used in the model as a validation set.

In short, we obtain the following samples:

1. Training Set (10,547 obs.)

2. Valid Set (3,649 obs.)

3. Normal Test Set (872 obs.)

4. Abnormal Test Set (872 obs.)

The final algorithm of the proposed method is shown in Algorithm 3.

Algorithm 3: Semi-supervised Variational Autoencoder Model

Data: Normal data-set x_N, Abnormal data-set $x_A(i)$ $i = 1, ..., N$, threshold $t^* \in T$

Split the Normal data-set: $(x_N^{train}(j), x_N^{valid}(k), x_N^{test}(i)) \in x_N$

Train VAE model using the data-set x_N^{train} and the validation data-set x_N^{valid}.

Initialize the encoder (f) and decoder (g) parameters.

for *e epochs to convergence of f, g parameters* **do**
  for *j to J* **do**
    We draw random samples L from $\epsilon \sim \mathcal{N}(0, 1)$;
    We get the random variable $z(j, l)$ from the deterministic transformation based on $\epsilon(j), x_N^{train}(j)$, $j \in J$, $l \in L$;
    We calculate $VAE_{loss}(z(j, l), x_N^{train}(j))$ and $VAE_{loss}(z, x_N^{valid}(k))$;
    Update encoder f, decoder g parameters;
  end
end

Use the trained model as anomaly detection model applied to $x^{test}(i) = \{x_N^{test}(i), x_A(i)\}$:

for $t \in T$ **do**
  for $i = 1$ *to* N **do**
    $z_\mu(i), z_{log(\sigma)}(i)$ from the encoder function f applied to $x^{test}(i)$;
    We draw random samples L from $z \sim \mathcal{N}(z_\mu(i), z_{log\sigma}(i))$ **for** $l = 1 to L$ **do**
      We get $\hat{x}_\mu(i, l), \hat{x}_{log(\sigma)}(i, l)$ to calculate the probability of the original data (parametrizing the decoder function g)
    end
    $reconstructionerror(i) = x^{test}(i) - \frac{1}{L}\sum_{l=1}^{L} p(x^{test}|\hat{x}_\mu(i, l), \hat{x}_{log(\sigma)}(i, l))$ where p is the likelihood of the x^{test} given the latent variables which derive the parameters of the original input variable distribution;
    if $reconstructionerror(i) > t$ **then** $x^{test}(i)$ is an anomaly, $\hat{y} = 1$;
    else $x^{test}(i)$ is not an anomaly, $\hat{y} = 0$;
  end
  We calculate $F_1(y, \hat{y})$
end

We get t^* from $F_1^* = argmax\{F_1(t)\}$.

5.6 Results

5.6.1 Misreporting Behavior: Testable Implications

The hypothesis to be tested in this section is whether all high risk individuals pretend to be low risk as basic insurance theory suggests, or if it is more nuanced in that only a subset misreports.

One limitation of this analysis is that the data on potential customers who got a negative simulated bonus, and therefore did not continue with the policy subscription, was not stored. This may affect the predictive power negatively, as we train with a lower proportion of sincere customers (assuming that the probability of a person telling the truth was higher in the rejected group). Also, this limits the descriptive analysis, because we can only hypothesize about risky potential customers between a limited range of the simulated bonus (from 0% to 60%). However, bonus below 35% indicates at least one self-declared claim in the last three years.

In Table 5.2 we present the lower bound distribution of the simulated bonus. Thus we have two potential risk simulated bonus outputs: 0% and 10%. Those results are reached by a combination of at least one year as insured, between one and three claims in the last five years or at least one claim in the last year. For instance, we have 89 potential customers that declared between one and three claims and a certain number of years as insured, but they finally got a negative bonus of 80% (because they had more claims than they declared).

Theory suggests that customers with the same type of risk will be pooled together. However, we find evidence that this relation is not empirically evident. The distribution is quite homogeneous between those who truly self-reported their nature (101 individuals reported their true risk) and those who misreported it (117 individuals). Additionally, 337 individuals reported a higher risk than the actual assigned risk. A feasible reason of this behavior is that they also revealed claims in which they were not the guilty drivers, as a precaution of future penalties. Contrarily, the database only reports cases in which they were considered as guilty drivers. The 438 cases (337 plus 101) that did reveal their true nature are not negligible at all if we consider that 872 individuals were finally labeled as risky customers.

One pending question is why people with similar risk chose to misreport or not. Ex-ante, they had not been informed that their answer would affect their outcome and they did not know that companies share this information[3]. However, they may infer that revealing certain information related with past claims could impact the final price. The reasons of misreporting seem to lie in unobservable differences

[3]It is important to remark that with the new General Data Protection Regulation implemented in European Union since 2018, this database is no longer available.

	Simulated (%)	
Adjusted (%)	**0%**	**10%**
-80%	49	40
-40%	0	0
-25%	7	1
-20%	7	1
-10%	5	3
-5%	3	0
0%	**22**	5
10%	**57**	**79**
20%	**0**	**0**
30%	**0**	**0**
35%	**44**	**27**
40%	**0**	**1**
45%	**88**	**32**
50%	**34**	**5**
55%	**40**	**9**

Table 5.2. Simulated and Adjusted Bonus for High Risk Potential Customers

among potential customers, such as different level of risk aversion or idiosyncrasy, that affect the choices of self-reporting (for a detailed analysis we recommend Cohen and Siegelman, 2010).

The question is, if there are any generalizable insights on what type of customers are likely to misreport. We examine a probit model with several variables typically associated with risk categorization as the independent variables, and using the probability of misreporting as the dependent variable. The information in Table 5.3 indicates that males or young people are more likely to misreport, which supports the findings of Dionne and Vanasse (1992). Additionally, we find that the number of license years reduce the risk and, surprisingly the vehicle age, which could be explained by a non-linear relation. Most important, however, is to see the effect of the cluster variables which are proxies between observables and the risk of internal customers. Results report that the four cluster variables are positively correlated with the risk of potential customers and, therefore, support the evidence that observable variables are correlated with the risk.

Bearing in mind that observable characteristics can explain risk, the main question we want to solve is whether or not self-reported data is useful to predict risk. Therefore, in the next section, we first present several models with only observable data. Then, we repeat the process, but this time including additional self-reported data, and compare results.

Pr(Risk)	dy/dx	Std. Error	Z
Gender	0.0287***	0.004	6.694
Age (18-30ys)	0.0229***	0.008	-2.858
License Years	-0.0024***	0.000	-15.14
Vehicle Age	-0.0016***	0.000	-5.698
ZIP Risk	0.0029***	0.001	2.868
Intermediary Risk	0.0162***	0.003	6.219
Vehicle Risk	0.041***	0.003	13.696
Customer Risk	0.008***	0.001	6.986

Table 5.3. Estimates Probit Model

5.6.2 Predicting Misreporting Behavior with Observable Characteristic

The input of the Variational Autoencoder model are the cluster variables, the potential insureds' observable characteristics and the answers to the questions related to their past performance. We predict whether a potential insured is risky before the policy is subscribed and compare the models' accuracy by using recall and precision metrics[4]. Prediction results for potential customers during the first six months of 2018 will be displayed in detail below.

Table 5.4 reports the results regarding the predictive performance of the variational autoencoder network. That is, the ability to predict who has lied during the underwriting process before subscribing the policy. This was applied to two test data sets (two homogeneous samples which comprise 50/50 normal/abnormal randomly selected cases). We compare these results with those obtained from the unsupervised outlier detection model Isolation Forest (Liu, et al. 2008), the supervised machine learning[5] Extremely Randomized Trees (Geurts, et al., 2006) and then with those of several deep-learning architectures: Deep Neural Networks (Cireşan et al., 2012), Deep Residual Connection (He et al., 2016) and Inception Model (Szegedy et al., 2016) and the Autoencoders model (Hinton and Salakhutdinov, 2006) also applied as a semi-supervised model. To test our results, we use the F1 Score which is a well-known accuracy test for binary classification problems[6].

While it is true that the observable characteristics have explicative power, they are not enough in predicting risk. None of the presented algorithms were able to separate risky individuals from non-risky individuals. For instance, both the AE

[4]Precision measures the ability to avoid including false positive cases – quality – and Recall measures the ability to capture true positive cases – quantity.

[5]For the supervising models (ERT, DNN, RC, IC), we use a stratified 5-Folds cross validation approach.

[6]$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$

Model	Sample 1			Sample 2		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Isolation Forest	0.616	0.103	0.177	0.654	0.078	0.139
Extremely Randomize Tree	0.674	0.355	0.465	0.651	0.321	0.430
Neural Network	0.500	1.000	0.667	0.500	1.000	0.667
Inception Model	0.546	0.486	0.515	0.529	0.477	0.502
Residual Connection Model	0.558	0.454	0.500	0.530	0.444	0.484
Autoencoder	0.500	1.000	0.667	0.497	0.984	0.660
Variational Autoencoder	0.500	1.000	0.667	0.500	1.000	0.667

Table 5.4. Comparative Results

and the VAE model classify all individuals as risky.

5.6.3 Combining Self-Reported Data and Observables

In Table 5.5, we report the results of combining the observable characteristics with the self-reported data about years as insured and past claims. Our hypothesis is data risk categorization will benefit from adding self-reported data.

Model	Sample 1			Sample 2		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Isolation Forest	0.909	0.252	0.395	0.898	0.264	0.408
Extremely Randomize Tree	0.966	0.603	0.743	0.959	0.594	0.733
Neural Network	0.500	1.000	0.667	0.500	1.000	0.667
Inception Model	0.927	0.174	0.293	0.938	0.174	0.294
Residual Connection Model	0.938	0.314	0.471	0.977	0.298	0.457
Autoencoder	0.735	0.800	0.766	0.721	0.837	0.775
Variational Autoencoder	**0.769**	**0.869**	**0.816**	**0.837**	**0.798**	**0.817**

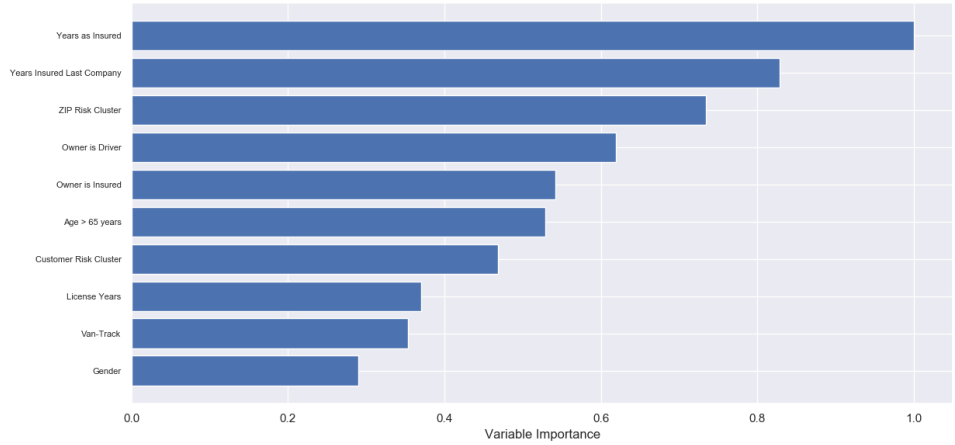
Table 5.5. Comparative Results

As a result, the predictive performance of VAE is significantly different from those of the other models. With a consistent 81.6%-81.7% F1-Score, the results show that the VAE model performs best in terms of predicting risky potential customers.

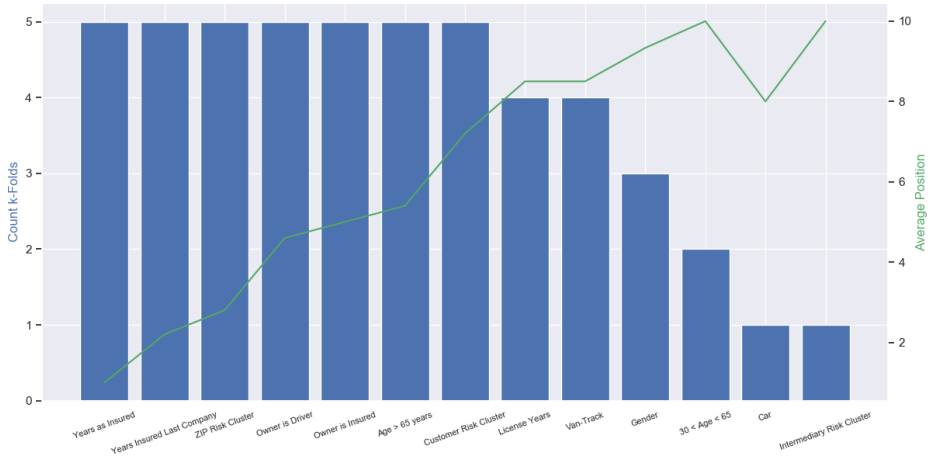
5.6.4 Feature Importance of Risk

A feature importance ranking is created to check which main variables drive risk. Feature importance provides a score that measures the importance of each attribute for improving performance (in terms of reducing the loss function). It is calculated as the improvement average for each attribute in every decision tree (when a splitting point is selected) weighted by the number of observations in the node. We present two plots in Figure 5.8. The first figure shows the ten most important

variables on average of a 5-Folds Cross Validation methodology, while the second shows, as a bar plot, the number of times a variable appears in the ranking, and as a line plot, its average position in the ranking.



(a) Average K-Folds CV

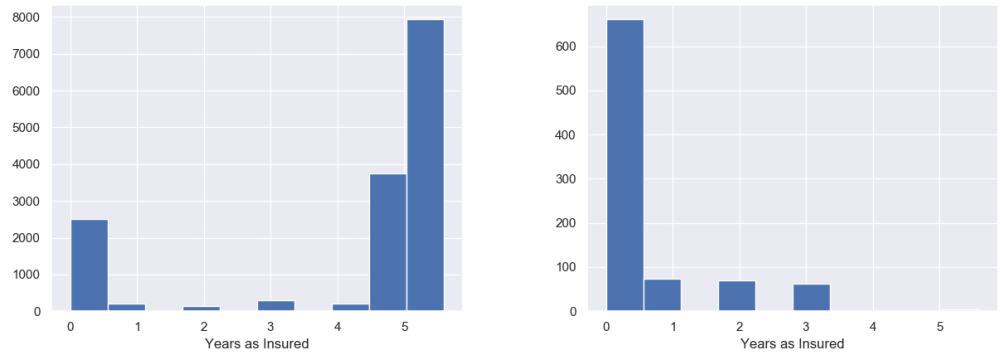


(b) Times and Average Position

Figure 5.8. Feature Importance Ranking

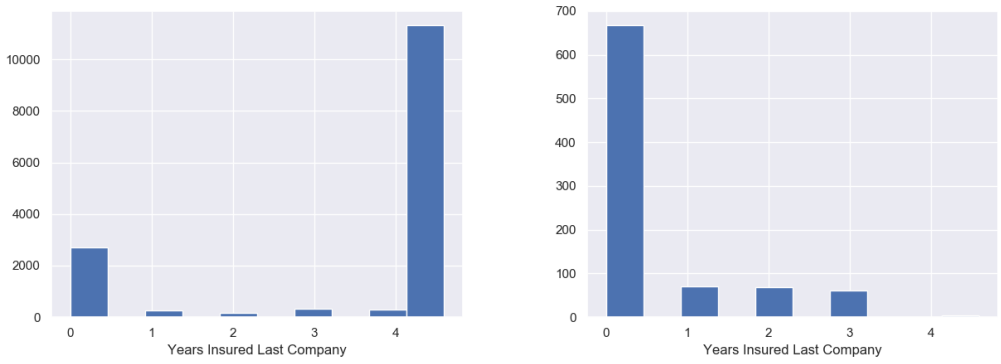
Surprisingly, previous self-reported claims does not emerge as a relevant variable. Rather, self-reported Years as Insured and Years Insured in the Last Company are consistently the most influential variables for predicting risky potential customers (with both always appearing in first and second place, respectively). In Figure 5.9 and Figure 5.10 we can observe this behavior more closely.

As can be seen, the correlation between risk and self-reported years as insured seems to present a negative pattern. The distributions clearly differ for normal and



(a) Normal Points (b) Abnormal Points

Figure 5.9. Years as Insured Correlation



(a) Normal Points (b) Abnormal Points

Figure 5.10. Years as Insured in the Last Company Correlation

abnormal points. Normal customers tend to self-report several years (more than 5 years) as insured, but also they tend to have had a long-term relationship (more than 5 years) with their last company. On the other hand, abnormal customers tend to report that they had a very short-term relationship with their last company, and none of them accredits more than 3 years as insured. Even more, the majority of the riskier customers self-reported less than one year as insured when, in fact, they had past claims.

A reasonable explanation could be that potential customers do not lie about years as insured, or do not have incentive to do so, because they do not associate it with a modification to the final bonus. Similarly, they are certainly likely to expect that their self-reported prior claims will be correlated with the final bonus offered and, therefore, have an incentive to underreport. On the other hand, customers with good records do not have any incentive to lie.

In line with this reasoning, why, therefore does the variable self-reported years as insured consistently explain risk and self-reported claims do not? On one hand, it seems that the answer lies in the fact that self-reported years as insured and self-reported years insured in the last company could implicitly reveal past behavior. Switching to a new company (after a very short term) may well conceal a strong motive, i.e. companies seek to deter risky customers from renewing their contract. As explained by dynamic learning theory (see Abbring et al., 2003a, 2003b; Hendel and Lizzeri, 2003; Finkelstein et al., 2005; Cohen, 2005; Finkelstein and McGarry, 2006), insurance companies may learn about the policyholders' risk type dynamically. In Spain, for instance, if a customer has a bad record, companies can choose not to renew the contract or they can deter renewal by imposing a high bonus. In short, a small number of years insured in the last company could be indicative of risky behavior. Inversely, this variable could be associated with a positive effect, indicating that those with more years as insured have a better record because they know how to behave. Therefore, customers with good records will reveal their true state. On the other hand, it could also be a natural tendency for bad record customers to underreport claims. But this behavior will naturally pool them with good record customers (because they have few or no claims to self-declare) and, in consequence, it could affect the ability of the algorithms to split normal from abnormal customers.

Two cluster variables are also found to be highly representative in the feature importance ranking: namely, the ZIP Risk Cluster and Customer Risk Cluster. As will be explained, they respectively associate a customer's ZIP code and certain observable characteristics with the weighted number of claims. However, we did not find any evidence that riskier intermediaries or vehicle characteristics had any relevance. We can, therefore, conclude that risky behavior is essentially related to

potential customer behavior.

Other relevant variables include whether the insured has signed the policy as the owner and whether he or she is also the first driver. A plausible explanation is that the insured as driver is likely to be more responsible with their own vehicle and with their own policy.

Likewise, age above 65, gender, and license years are also important variables. However, one of the main problems with feature importance ranking is that we do not know the direction or magnitude of the effects. We cannot therefore make any subjective assumptions about age or gender, but we would expect license years to be negatively correlated with risky customers (see Figure 5.24).

Finally, the last variable among those ranked in the top ten is the only one related to vehicle characteristics: If it is a van or truck. This could reflect the fact that, in general, they are more likely to be used as working vehicles and, so, of being on the road for many more hours, which makes them more likely to be involved in accidents and to deteriorate.

5.7 Conclusion

During the underwriting process, insurance companies face a severe problem of asymmetric information, i.e. they cannot distinguish between different risk behaviors. This information asymmetry is essentially attributable to the existence of unobservable characteristics and a group of customers that might underreport their prior claims. As a result, an unfair price is offered to sincere customers and prices become noncompetitive for the company.

In light of these challenges, we have sought to address the following question: Are all potential customers misreporting their true nature as insurance theory suggests? Our first main finding is empirical evidence against this situation, which leads to the following question: Can we combine this information with the traditional risk categorization methods to predict risky potential customers before an insurance policy is signed (and, in the end, offer a fair price)?

Insurance markets are an ideal setting in which to conduct empirical analyses. They dispose of large, rich databases about their own customers, and moreover, they lend themselves to the use of machine-learning and deep-learning models, which can exploit their large data-sets to find complex patterns. Here, the empirical evidence provided by the Spanish market is especially appropriate for evaluating our main objective, given that it operates with an external service via which leading companies share historical customer information. Thus, we can validate if we are actually able to predict risk based solely on customers' observable characteristics,

and the potential benefits of combining this with self-reported data.

Here, we have proposed using cluster variables based on the internal customers' risk as a proxy for the risk of potential customers. By matching the observable characteristics of the two groups, we can approach the unknown risk. Using K-means++ and a variety of validation metrics, we have created several risk rankings by customers, vehicles, ZIP codes and intermediaries. Next, we have used these cluster variables – plus the observable variables – as input for a deep variational autoencoder model. Here, one major computational issue is the fact that insurance companies dispose of a considerable amount of data that may or may not be related to the risk of a potential customer. Moreover, as we have shown, only a small number of customers fail to reveal their true risk during the underwriting process (5.5%). The VAE model allows us to reduce the data to its true nature and, at the same time, the variational autoencoder can be transformed into a powerful outlier model.

We have used a real-world data-set provided by a leading insurance company and have shown how the model works in comparison with other machine learning and deep learning models. The second main contribution we identify is the ability to detect potential new customers who fail to reveal their risky nature by combining self-reported data and observable characteristics. Our empirical evaluation supports the hypothesis that the VAE model outperforms other benchmark techniques, achieving a precision of 77-84% and a recall of 80-87%.

The third major contribution of the methodology is its ability to identify the most relevant variables for predicting risk. Here, we have shown that years as insured has a significant influence while self-declared prior claims are unimportant. This may be explained by the lower incentives to misreport associated with this variable and its high correlation with customer performance. Additionally, two of the cluster risk variables were found to be especially significant: one related to customer characteristics and the other to zip code. The following were also found to be systematically important variables: if the insured was the owner and first driver in the policy, if the customer's age was higher than 65, if the insured was male or female and, finally, the number of license years.

Finally, a theoretical model was presented which supports the fact that the combination of self-reported data and risk categorization can lead to an equilibrium in which the monopoly gets higher profits than the private information equilibrium, and that it can even reach a public information equilibrium.

5.8 Appendix

5.8.1 Clustering Observable Risk Variables

To construct the clusters, we use K-means++ (Arthur and Vassilvitskii, 2007). K-means (Lloyd, 1982) is a well-known distance based algorithm that operates by choosing random centers. It tunes the centers' location by minimizing the sum of squared Euclidean distance from the points to the center and then assigns each point to a particular center. The assignment of a point to a unique center provides the cluster composition. The main problem with K-means is that it is highly sensitive to initialization, i.e., the loss function is very susceptible to local minima. Poor randomization of the initial centroids or seeds will therefore result in suboptimal clustering. In contrast, K-means++ uses a robust initialization mechanism that guarantees convergence to an optimal solution. The idea is to maximize the distance between initial centers and so create new centers. These are randomly drawn from the probability distribution which emerges from the comparison between n possible centers.

As this is an unsupervised algorithm, performance evaluation is typically more complex as we do not have a defined target variable. Here, we follow a standard procedure: First, we evaluate cluster existence and, second, we internally validate the clusters.

To evaluate whether a cluster exists, we compare the hypothesis of the pattern's existence with the hypothesis of a uniformly distributed data-set. Here, we use the Hopkins statistic (Hopkins and Skellam, 1954; Banerjee and Dave, 2004) to evaluate the null hypothesis that the data are generated by a Poisson process (homogeneous distribution). The statistic takes values between 0 and 1, where a value close to 1 indicates that the data are highly clustered, 0.5 indicates they are random and a value close to 0 indicates they are uniformly distributed.

$$H = \frac{\sum_{i=i\ldots n} d(p_i)}{\sum_{i=1\ldots n} d(p_i) + \sum_{i=1\ldots n} d(q_i)}$$

where $d(p_i)$ is the distance of the point p_i to its nearest neighbor. $d(q_i)$ is the distance of the point q_i to its nearest neighbor, and q_i are n random points uniformly distributed in the same space of the data-set. The final Hopkins statistic result is derived from computing the H index and calculating the average.

Internal validation is a way of validating clusters when there are no labels available and is based on cluster compactness (i.e., variance or distance, how closely related the points are) and separation (i.e., distance or density, how well-separated a cluster is from other clusters). A wide range of internal validation indexes exist but

here we focus on three measures that have been shown to perform well in a wide range of situations (Saitta et al., 2007; Liu et al., 2010; Arbelaitz et al., 2013): The Silhouette score, the Calinski-Harabasz score and the Davies-Bouldin index.

The Silhouette score or SC (Rousseeuw, 1987) evaluates the pairwise difference between and within the cluster distances. It also obtains an optimal number of clusters and is defined as:

$$SC_i = \frac{b-a}{max(a,b)}$$

where a is the average distance between a sample and the remaining points of the same class; and b is the average distance between a sample and the remaining points of the nearest cluster. If it tends to -1 we have an incorrect cluster definition, when it is 0 we have overlapping clusters, and when it tends to 1 we have compact and separate clusters. The disadvantage of this metric is that it is higher for convex clusters than for density-based clusters.

The Calinski-Harabasz score or CH (Calinski and Harabasz, 1974) evaluates the average sum of squared dispersion between and within clusters, and it is defined as:

$$CH = \frac{Tr(B_k)N-k}{Tr(W_k)k-1}$$

where $B_k = \sum_q n_q(c_q-c)(c_q-c)^T$ is the dispersion between clusters and $W_k = \sum_{q=1}^{k}\sum_{x\in C_q}(x-c_q)(x-c_q)^T$ is the dispersion within clusters. For the cluster k, CH is the ratio of the average dispersion between clusters and the dispersion within. The disadvantage is that it tends to be higher for convex clusters.

The Davies-Bouldin index or DB (Davies and Bouldin, 1979) evaluates, for each cluster, the similarity between that cluster and the remaining clusters, and is defined as:

$$DB = \frac{1}{n}\sum_{i=1\ldots n} max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$$

where n is the number of clusters, c_i the cluster i centroid, σ_i the average distance between the points in i and the centroid, and $d(c_i,c_j)$ is the distance between the centroids i and j.

Based on these three metrics, we use the k-means++ approach for four different categories of clusters: 1) ZIP code, 2) Intermediary, 3) Object and 4) Customer.

For ZIP Code risk, we use a very simple approach. We group postal codes by taking the weighted sum of vehicle claims by the number of policies.

For Intermediary risk, as the intermediary can sell other products, we create two additional variables: The vehicle intermediary risk (the number of vehicle claims

over the number of vehicle policies) and the vehicle policy share (the number of vehicle policies sold by the intermediary and the number of total policies). We then calculate the clusters grouped by the identification number of the intermediaries and normalized by their antiquity.

For Object, we create a range of values based on vehicle price. Then, we group the vehicles by this price range, category, usage, type, number of years and whether it is a heavy vehicle (more than 3500 kg).

For Customer, we group the number of claims and policies (normalized by the time as active customer) by age, license years, risk license (less than a year), second driver risk license (less than a year), risk driver age (between 18-21 years), risk second driver age (between 18-21 years), gender and nationality.

Table 5.6 lists the results.

Risk	**k-clusters**	**Existence**	**Internal Validation**		
		H	**SC**	**CH**	**DB**
Postal Code	6	0.939	0.624	7,781	0.536
Intermediary	4	0.998	0.738	2,232	0.483
Object	3	0.997	0.840	37,869	0.490
Customer	6	0.998	0.780	1.992	0.358

Table 5.6. Cluster Validation using K-Means++

Based on these results, we obtain acceptable values to infer cluster existence as well as the compactness and separation of these clusters.

ZIP code

We depict the ZIP codes in Figure 5.11. Clusters are ranked from 0 to 5, where 0 is the cluster with the minimum weighted claims average, and 5 the maximum weighted average.

As expected, the company's customers concentrate in the main urban areas. Although the risk seems to be more concentrated in the south, the riskiest places correspond to two postal codes in Bilbao (with more than two claims per policy per year, when, on average, there are 0.30 claims per policy per year).

Intermediary

Intermediaries provide us with four clusters. The lowest cluster has an average of 1.37 claims per policy sold per year, and the highest has an average of 2.52 claims per policy sold per year. In fact, the latter corresponds to a single intermediary who while being a leading seller performs poorly.

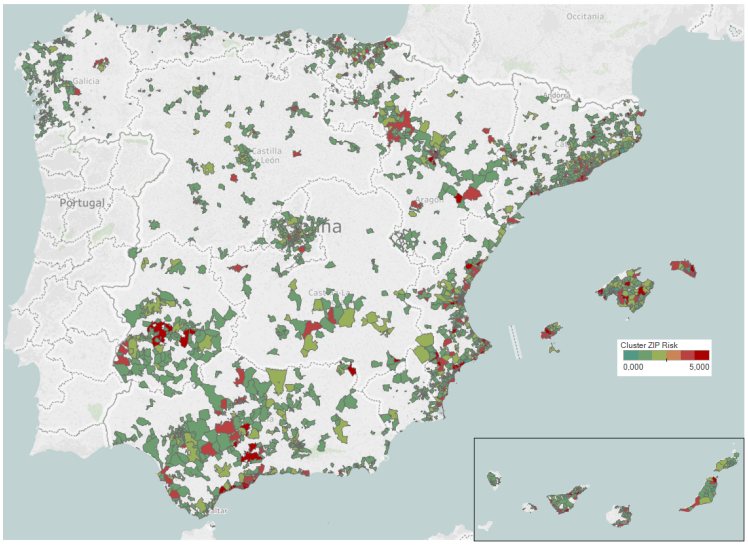


Figure 5.11. ZIP Risk Clusters Map

Object

We have just three vehicle clusters. In Appendix 5.8.4, we show a set of different plots with the vehicle variables used. For example, in Figure 5.19, we can see that the majority of insured vehicles are owned by private individuals, and that the distribution is concentrated between 0 and 2 claims per policy per year. Figure 5.20 plots the main categories of vehicle types. Cars have more claims per policy on average than motorbikes and present a very similar distribution to that of Vans or Trucks. In Figure 5.21 we plot the claims by vehicle value. The tails seem to reflect lower risk levels than the average values. Finally, in Figure 5.22 we see an abrupt fall in risk after 30 years, possibly correlated with collectible classic cars.

Customer

In Figure 5.23 we can see that older customers actually present a lower and decreasing risk ratio. As expected, customers with new licenses present a significant risk difference compared to customers with more than five years' driving experience (Figure 5.24). Neither gender nor nationality appear to be related to higher risk ratios (Figure 5.25 and Figure 5.26).

5.8.2 Variational Autoencoder Model Validation

Theoretical Definition

When using Variational Autoencoder models (VAE) we are interested in finding a joint distribution between input x and the latent variables z, which are not part of

the data-set but that are representations of certain properties of the input. Using a variational inference model (an encoder $Q(z|x)$), we can approximate the input data and their attributes. First, an encoder network turns the input samples x into two parameters in a latent space, which we note as z_μ and $z_{log(\sigma)}$ (we assume the encoder to be a multivariate Gaussian). The inference model generates the latent vector z from input x.

Second, we randomly sample similar points z from the latent normal distribution[7] that is assumed to generate the data via $z = z_\mu + e^{z_{log(\sigma)}/2} * \epsilon$, where ϵ is a random normal tensor. This is known as a reparametrization trick. The decoder takes samples from z to reconstruct x. But, backpropagation cannot pass through a stochastic layer. Therefore, we take the sampling process outside to be computed as $z = z_\mu + e^{z_{log(\sigma)}/2} * \epsilon$. This overcomes the problems with high variance, and avoids the random variable from the original distribution (Paisley et al., 2012). This reparametrization should ensure that z follows the distribution of the encoder. Finally, a decoder network $P(z|x)$ maps these latent space points z back to the original input data x.

However, to estimate the true distribution of our inputs, we must identify the relationship between the encoder and the decoder. To do so, we can use the Kullback-Leibler divergence (KL) to obtain the distance between these two conditional densities.

$$D_{KL}(Q(z|x)||P(z|x)) = E[logQ(z|x) - logP(z|x)]$$

Using Bayes' theorem and rearranging terms, this can be defined as:

$$logP(x) - D_{KL}(Q(z|x)||P(z|x)) = E(logP(x|z)) - D_{KL}(Q(z|x)||P(z))$$

On the left-hand side we have $P(x)$, the true distribution of x, and $D_{KL}(Q(z|x)||P(z|x))$, the error due to the distance between the encoder and the decoder. If we minimize the KL distance, we are better able to encode the attributes x to z. The right-hand side indicates that we have to maximize $E(logP(x|z))$ (the decoder seeks to reconstruct x based on z samples – the Reconstruction Loss) and minimize the KL distance between the encoder Q and the prior P. By assuming that $Q(z|x)$ is a multivariate Gaussian, the right KL term can be simplified to:

$$-D_{KL}(Q(z|x)||P(z)) = \frac{1}{2}\sum_{j=1}^{J}(1 + log(\sigma_j)^2 - (\mu_j)^2 - (\sigma_j)^2)$$

[7] We assume that the relationship between the variables in the latent space is much simpler than that in the input space.

with $\sigma_j \to 1$ and $\mu_j \to 0$. We can summarize the VAE function as:

$$\zeta_{VAE} = \zeta_{RC} + \zeta_{KL}$$

The model's parameters are trained via these two loss functions: the reconstruction loss forcing the decoded samples to match the initial inputs (exactly as in autoencoder networks) - ζ_{RC}, and the divergence between the learned latent distribution and the prior distribution acting as a regularization term -ζ_{KL}.

Model Setup

The first step is to construct the encoder network. As previously stated, this takes the input vector and calculates the mean (z_μ) and the log variance ($z_{log(\sigma)}$) of the Gaussian distribution. Using both, we create a latent variable z by randomly sampling points from the latent normal distribution, which is assumed to be generated by $z = z_\mu + e^{z_{log(\sigma)/2}} * \epsilon$ (where ϵ is a random normal tensor with mean 0 and standard deviation 1). We then create the decoder network symmetrically with that of the encoder. Next, we train the model using the variational autoencoder. We set the loss function to be the sum of the reconstruction loss (based on the autoencoder loss) and the KL divergence regularization function[8] between the learned distribution and the prior distribution (the latent loss):

$$VAE_{loss} \; = \; AE_{LOSS} + 0.5 * (1 + z^2_{log(\sigma)} - z^2_\mu - e^{z_{\log(\sigma)}})$$

The left-hand side is the reconstruction error (or the autoencoder loss) and is defined as the difference between the input vector x and the reconstruction z. The right-hand side is the latent loss (ζ_{KL}).

To calculate the reconstruction error, we use the mean squared error between the input nodes x and the decoded values $\hat{x}$. This calculates the squared difference between predicted values ($\hat{x}$) and the actual value x. The advantage of using mean squared error in this scenario is that it is more sensitive to outliers. If we understand that outliers represent anomalies, marked differences between input and predicted data should be highlighted by a greater reconstruction loss.

In the experiments, we train autoencoder and variational autoencoder architectures using symmetrical decoder/encoder designs. Figure 5.27 illustrates the optimized deep-learning architecture used.

We use dense layers, which means that every neuron in the layer is fully connected to the next one. There are no absolute rules for choosing the number of

[8] Kullback-Leibler divergence is defined between two distributions P, Q as $D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$.

layers and neurons. On the one hand, using too few neurons can lead to underfitting but, on the other, using too many may result in overfitting. Similarly, if we use more layers, the model can learn more complex representations, although this may result in a loss of information as the data become compressed.

To avoid overfitting, we used three different regularization methods: a dropout of 30%, sparsity constraints, and early stopping set to 2 epochs. The dropout is a regularization method that randomly drops out a certain number (or percentage) of layer outputs. It experiments with different architectures of the same network in parallel and the robustness of this model increases as the training process is noisy and we force the nodes to actually learn a sparse representation (Srivastava et al., 2014). Likewise, sparsity constraints restrict some nodes from being triggered, as some of them are restricted or not equal to zero. We use both L1 and L2 regularization functions. L1 adds a restriction to the nodes as the sum of the square of the weights, and L2 adds a restriction as the sum of the absolute value of the weights. Finally, we use early stopping which provides a rule of how many iterations we can run before the learner overfits (i.e., when the test loss starts to be worse than the training loss).

We use an Adam optimizer (Kingma and Ba, 2015) as this has several advantages over the classic stochastic gradient descent. The objective with optimization is to minimize the loss function, i.e., the way in which we trace the curve of the loss function to its minima. Instead of using a single learning rate, it adapts the parameter calculating an exponential moving average of the gradient and the squared gradient, controlling the decay rates of these moving averages. As Ruder (2016) shows, the Adam optimizer works better in empirical problems than other optimization methods. Its main advantages are that it improves performance on problems with sparse gradients and it performs well on non-stationary problems.

The activation functions define the output of the neurons given the previous input. Each autoencoder was trained using a hyperbolic tangent activation function[9] which has the advantage of being less likely to become stuck (as it strongly maps negative inputs to negatives outputs). The output is in the range (-1, 1), therefore, it is recommended to re-scale the data within this range. The final activation function is a sigmoid whose output values lie between 0 and 1, and is therefore ideal for binary problems.

In our experiment, we use 1000 epochs (i.e., the number of times we go through the training set) and 20 steps per epoch (i.e., the batch size definition). The algorithm of the proposed method is shown in algorithm 3.

As An and Cho (2015) explain, the main advantages of using VAE models as

[9] $\frac{e^z - e^{-z}}{e^z + e^{-z}}$

opposed to autoencoders (essentially because of their deterministic nature) can be summarized as follows: First, latent variables are derived from the probabilistic encoder, facilitating the use of the variance of the latent space. By so doing, we can exploit the variance differences between normal and abnormal data (we expect abnormal data to present greater variance and to have higher reconstruction error than normal data). Second, the reconstruction error considers not only the differences between input and decoded values, but also the variance of the reconstruction (by considering the variance of the distribution). Variables that have a large variance will tolerate larger differences between the reconstruction and the original input (and inversely with a small variance).

Reconstruction Error Validation

Our model is not a typical predictive model and, as such, in order to predict anomalies (i.e., potential customers that did not tell the truth about their past performance), we do not use probabilities as in the traditional case. What we seek to do is to reconstruct the input values and compare them with the original input vector. We would expect anomalies to have a greater error than normal points (and we use the mean squared error to compare these differences). If the error is higher than a defined threshold, then the points are considered anomalies. Here, we expect the reconstruction of the VAE error to encounter problems when reconstructing anomalous cases. Finally, we can check the validity of these results by comparing them to the real target value.

After training with only normal cases and using the validation-set to optimize the model, we are able to check how the errors behave, as seen in Figure 5.12. Both, the training and the valid error seem to converge at approximately 40 epochs. The question however is if the error loss is low enough.

We use the VAE model to predict the normal test data and the abnormal test data, and then we calculate the error differences between their real and predicted values. From Figure 5.13, the differences between the normal and the abnormal reconstruction error distributions can be appreciated.

In Figure 5.14 we plot the reconstruction error of both test samples used (872 normal cases and 872 abnormal cases). The differences between the normal and abnormal points in the VAE reconstruction are clear. The abnormal points have a higher reconstruction error than that of the normal points. By using an optimized threshold, we can almost separate both classes, which illustrate the power of the VAE model as an outlier algorithm.

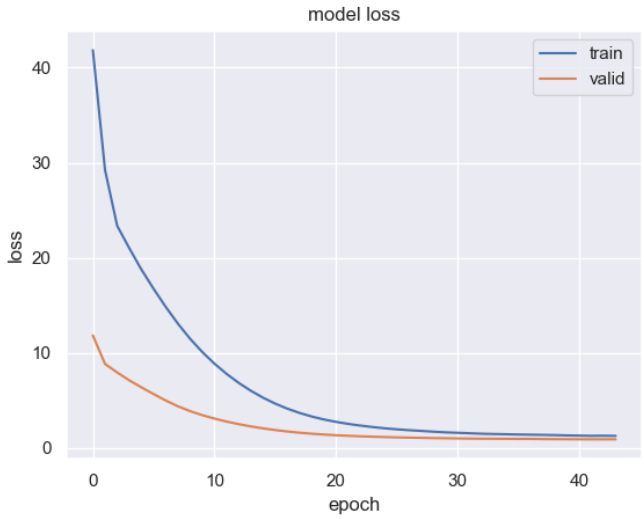
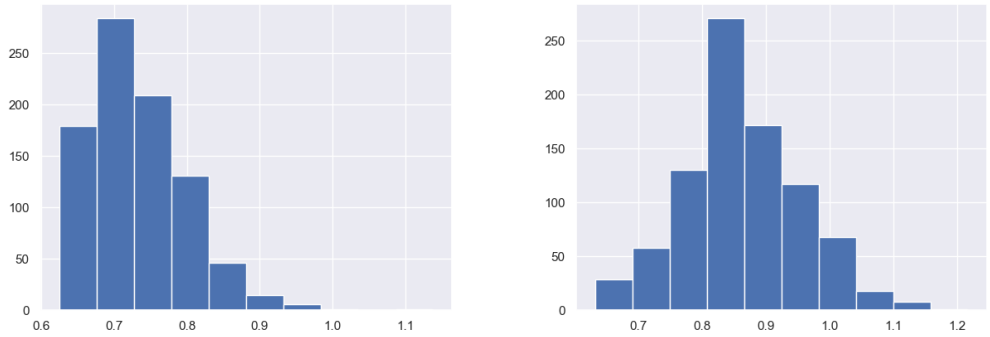


Figure 5.12. Error Convergence



(a) Reconstruction Normal Error

(b) Reconstruction Abnormal Error

Figure 5.13. Difference between True and Predicted Values

(a) Reconstruction Error Sample 1

(b) Reconstruction Error Sample 2

Figure 5.14. Reconstruction Error

5.8.3 Internal Cluster Validation Plots

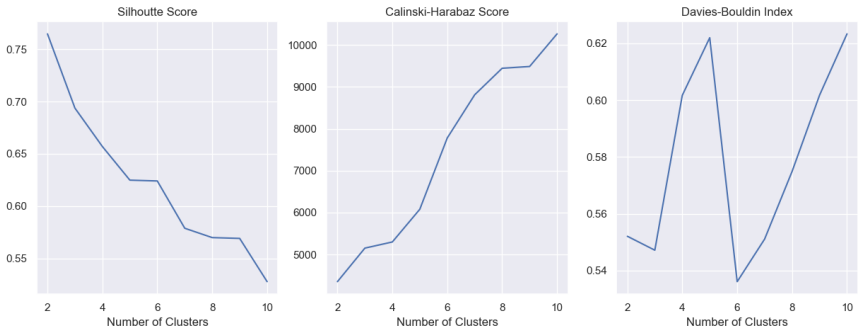


Figure 5.15. Cluster Internal Validation: Postal Code

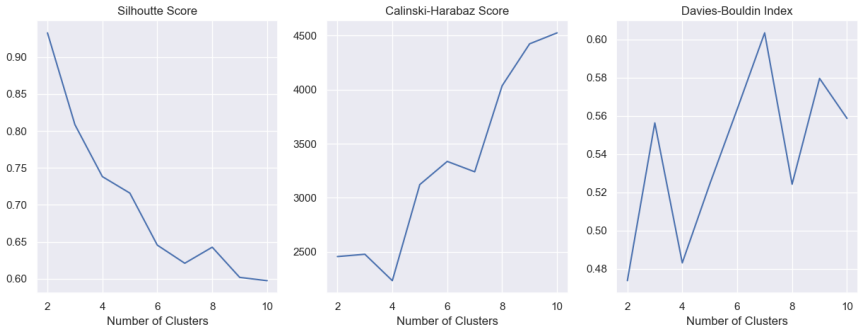


Figure 5.16. Cluster Internal Validation: Intermediaries

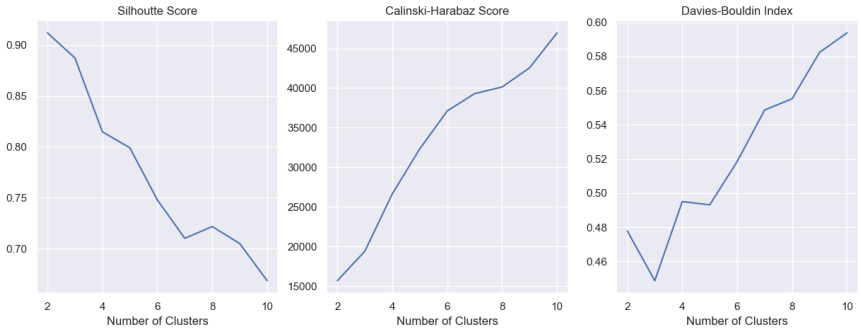


Figure 5.17. Cluster Internal Validation: Object

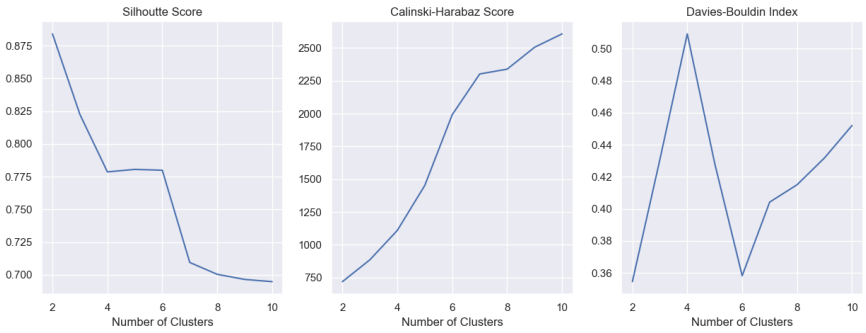


Figure 5.18. Cluster Internal Validation: Customer

5.8.4 Cluster Statistics Plots

Vehicle

Figure 5.19. Vehicle Usage versus Claims

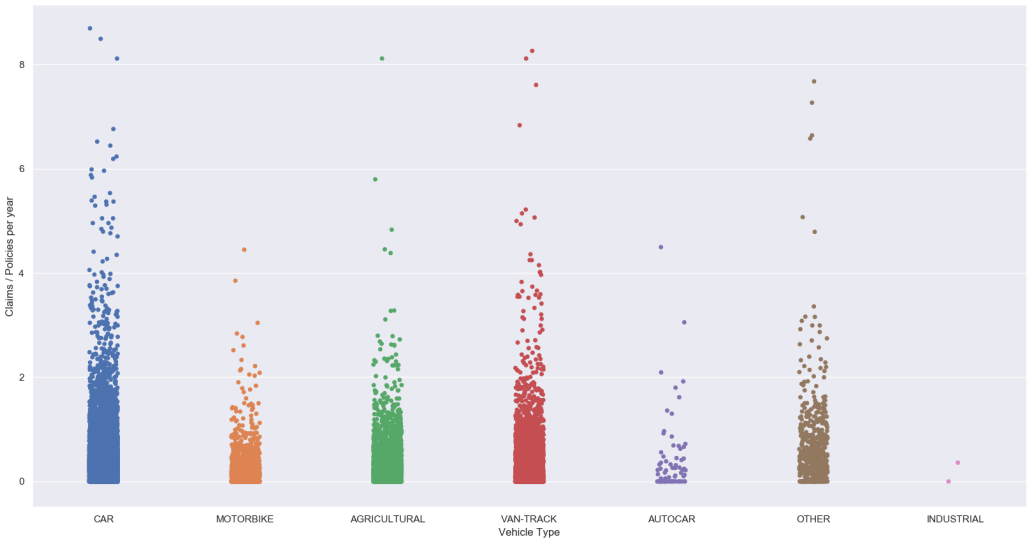


Figure 5.20. Vehicle Type versus Claims

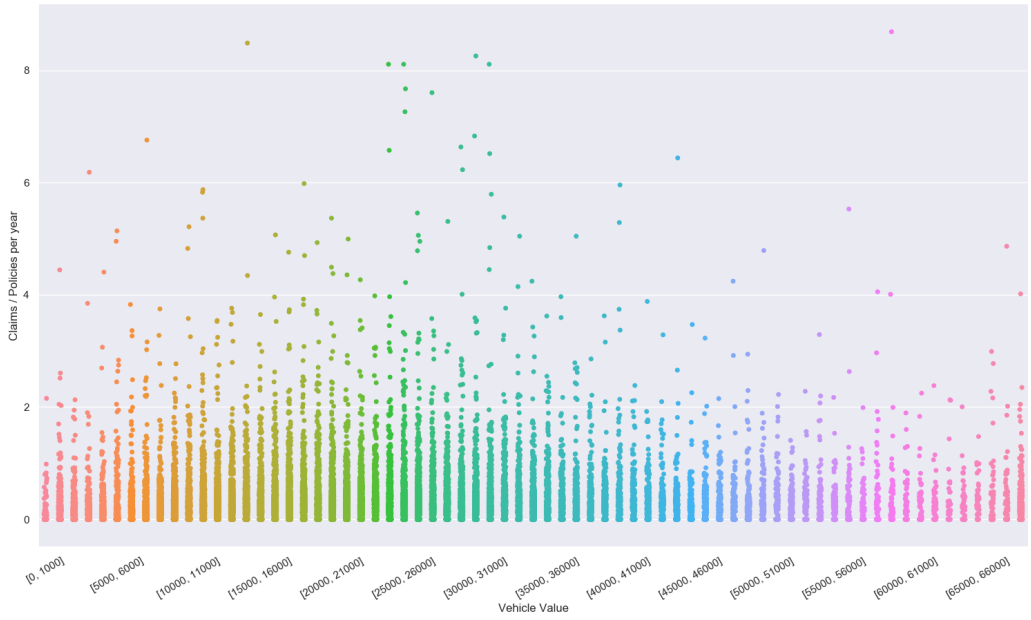


Figure 5.21. Vehicle Value versus Claims

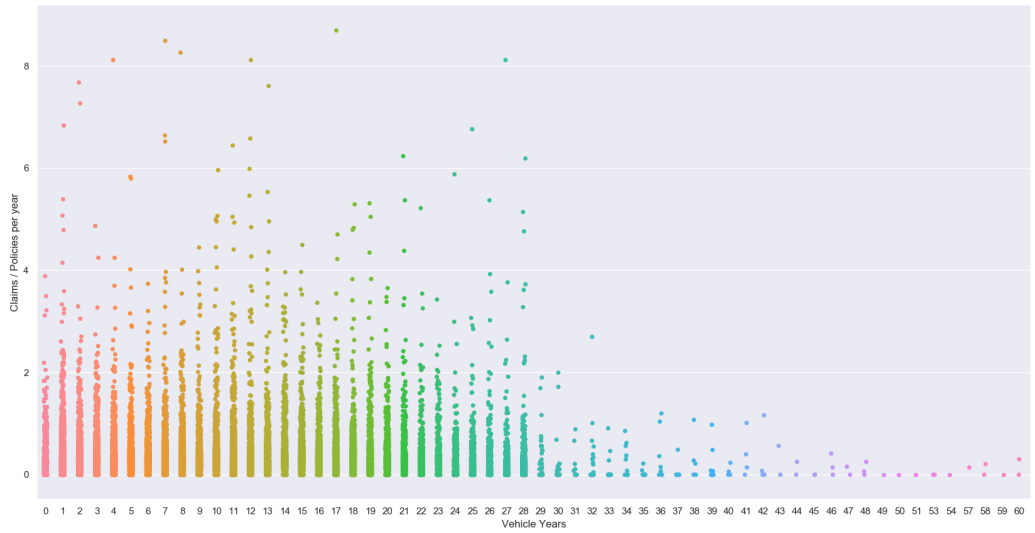


Figure 5.22. Years of the Vehicle versus Claims

Customer

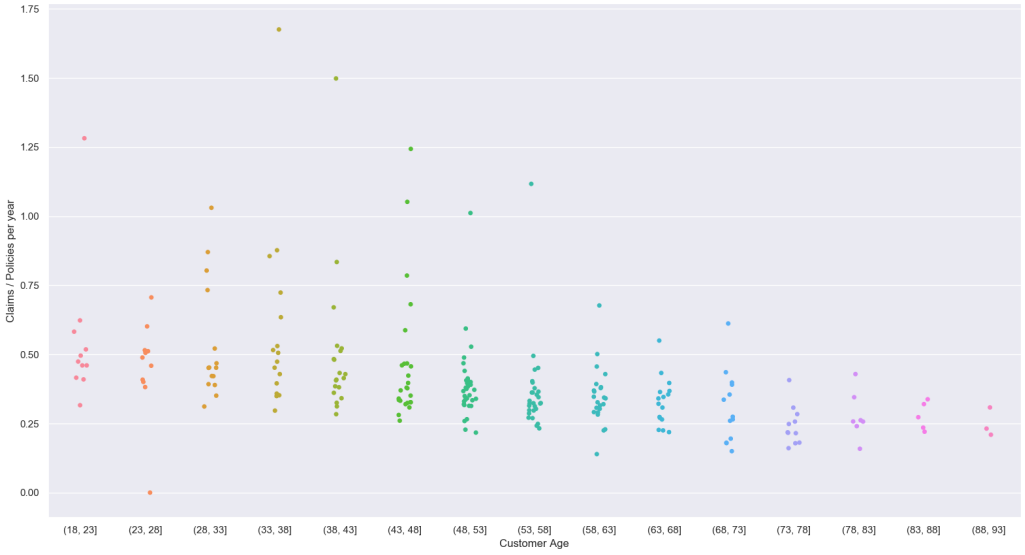


Figure 5.23. Customer Age versus Claims

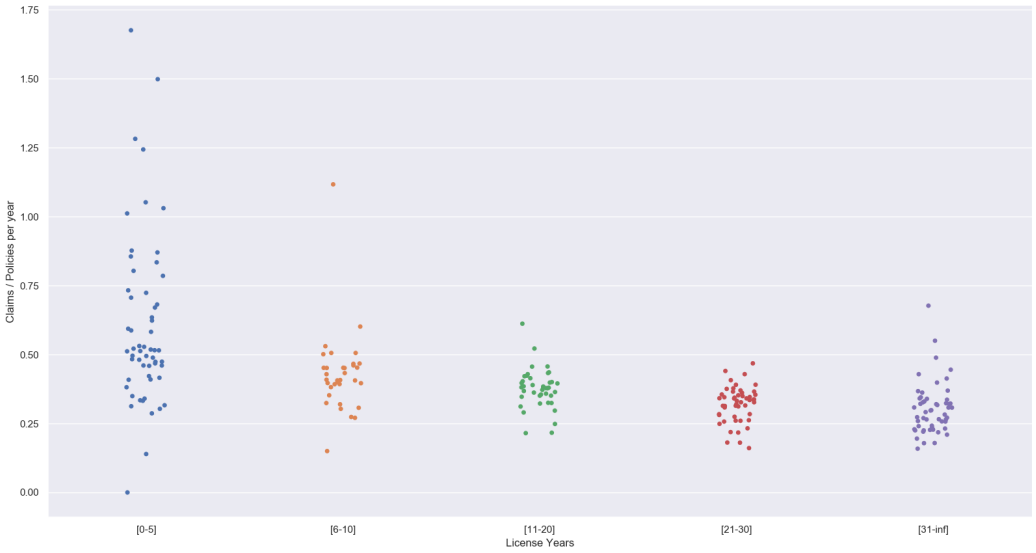


Figure 5.24. License Years versus Claims

Figure 5.25. Gender versus Claims

Figure 5.26. Whether Spanish or foreigner customer versus Claims

5.8.5 Network Architecture

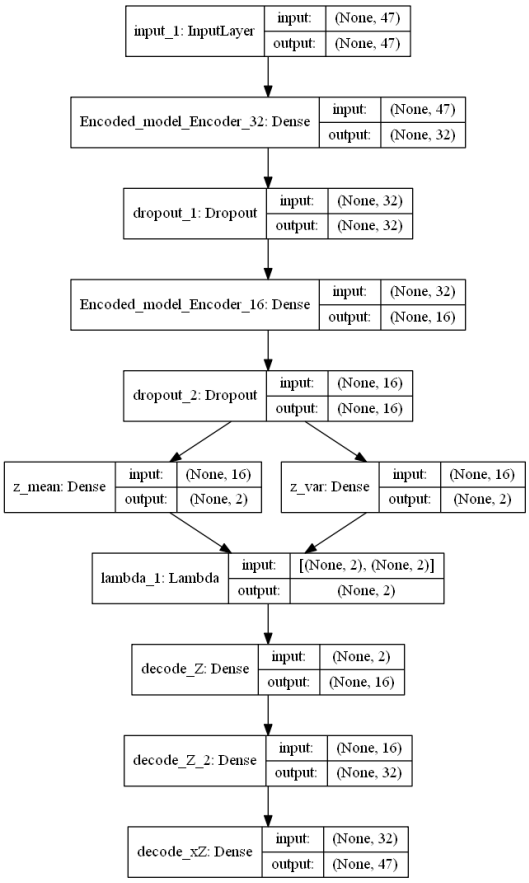


Figure 5.27. Variational Deep Autoencoder

6 Conclusions

The presented discourse followed several topics where every new chapter introduced an economic prediction problem and showed how traditional approaches can be complemented with new techniques like machine learning and deep learning. These powerful tools combined with principles of economic theory is highly increasing the scope for empiricists. Chapter 3 addressed this discussion. By progressively moving from Ordinary Least Squares, Penalized Linear Regressions and Binary Trees to advanced ensemble trees. Results showed that ML algorithms significantly outperform statistical models in terms of predictive accuracy. Specifically, ML models perform 49-100% better than unbiased methods. However, we cannot rely on parameter estimations. For example, Chapter 4 introduced a net prediction problem regarding fraudulent property claims in insurance. Despite the fact that we got extraordinary results in terms of predictive power, the complexity of the problem restricted us from getting behavioral insight. Contrarily, statistical models are easily interpretable. Coefficients give us the sign, the magnitude and the statistical significance. We can learn behavior from marginal impacts and elasticities. Chapter 5 analyzed another prediction problem in the insurance market, particularly, how the combination of self-reported data and risk categorization could improve the detection of risky potential customers in insurance markets. Results were also quite impressive in terms of prediction, but again, we did not know anything about the direction or the magnitude of the features. However, by using a Probit model, we showed the benefits of combining statistic models with ML-DL models. The Probit model let us get generalizable insights on what type of customers are likely to misreport, enhancing our results. Likewise, Chapter 2 is a clear example of how causal inference can benefit from ML and DL methods. These techniques allowed us to capture that 70 days before each auction there were abnormal behaviors in daily prices. By doing so, we could apply a solid statistical model and we could estimate precisely what the net effect of the mandated auctions in Spain was. This thesis aims at combining advantages of both methodologies, machine learning and econometrics, boosting their strengths and attenuating their weaknesses. Thus, we used ML and statistical methods side by side, exploring predictive performance and interpretability.

Several conditions can be inferred from the nature of both approaches. First,

as we have observed throughout the chapters, ML and traditional econometric approaches solve fundamentally different problems. We use ML and DL techniques to predict, not in terms of traditional forecast, but making our models generalizable to unseen data. On the other hand, traditional econometrics has been focused on causal inference and parameter estimation. Therefore, ML is not replacing traditional techniques, but rather complementing them. Second, ML methods focus in out-of-sample data instead of in-sample data, while statistical models typically focus on goodness-of-fit. It is then not surprising that ML techniques consistently outperformed traditional techniques in terms of predictive accuracy. The cost is then biased estimators. Third, the tradition in economics has been to choose a unique model based on theoretical principles and to fit the full dataset on it and, in consequence, obtaining unbiased estimators and their respective confidence intervals. On the other hand, ML relies on data driven selection models, and does not consider causal inference. Instead of manually choosing the covariates, the functional form is determined by the data. This also translates to the main weakness of ML, which is the lack of inference of the underlying data-generating process. I.e. we cannot derive economically meaningful conclusions from the coefficients. Focusing on out-of-sample performance comes at the expense of the ability to infer causal effects, due to the lack of standard errors on the coefficients. Therefore, predictors are typically biased, and estimators may not be normally distributed. Thus, we can conclude that in terms of out-sample performance it is hard to compete against ML models. However, ML cannot contend with the powerful insights that the causal inference analysis gives us, which allow us not only to get the most important variables and their magnitude but also the ability to understand economic behaviors.

Thereby, to start with this thesis we tried to answer the question to what extent liberalized electricity markets react to regulations that attempt to increase competition and reduce price volatility. Respectively, in Chapter 2, taking Spain's experience as a framework for empirical analysis, we have examined the impact of mandated auctions in daily electricity prices. Particularly, we tried to predict collusive patterns when fixed-price forward contracts were applied in the Spanish electricity market.

The first target was to demonstrate when the abnormal price phases occurred. We used an ARMAX model combined with instrumental variables to reproduce the price dynamics in order to control for autocorrelation and endogeneity. By doing so, we found that 70 days before the mandated auctions, prices were significantly higher than other periods. This time window was also validated in two ways: Firstly, with a falsification test by estimating the same model in a year in which the policy was canceled. Secondly, by using a Long Short-Term memory network as an anomaly detection model.

These results led directly to another question, namely, what the economic impact

of mandated auctions in prices was. We employed a triple differences estimation, where the goal of the empirical exercise was to capture the effect of fixed price long-term auctions held in 2013 and which impacted Spain. To control for economic changes that were unrelated to the program, we used a market which was unaffected by the regulation (Nord Pool market) and a year in which regulation was canceled (2014).

Our analysis suggests that prices increased by 15 percent 70 days before the mandated auctions (compared to prices in a control group, here the Nord Pool market) and that this effect disappeared once the auctions were no longer held.

The first conclusion of this analysis is that results present evidence contrary to the literature to date which argues that fixed-price forward contract obligations increase competition and approximates prices to the marginal cost.

Second, based on the theoretical model developed, we conclude that two main factors could lead to this result: On one hand, preexisting natural concentration in the Spanish electricity market serves as an incentive to avoid pro-competitive regulations. On the other hand, fixed tariff in a market characterized by high volatility induces firms to charge a risk premium.

Finally, the presented results, naturally, lead to policy recommendations. Though, the regulation had the intention to stabilize the consumers' tariff cost, it did not take into account the specific characteristics of the Spanish electricity market. The reaction power of the Spanish firms seems to stem from high concentration (64% of the generation capacity was in hand of two firms) and a pivotal index rate below the threshold recommended by the European Commission. In addition, a low level of interconnectivity could also have contributed to a collusion environment. Moreover, liberalization exposed the electricity market to high price volatility oscillating between 0 and 180€/MHw in the period of analysis. Thereby, policy makers need to keep in mind, when designing regulatory policies such auctions, that they have to take into account the inherent characteristics of these markets. A format of repeated auctions and fixed prices in an environment of natural concentration and high volatility in prices will logically lead to a noncompetitive reaction.

Chapter 3 evaluated the potential of smart card data to predict public transport demand. By using a smart card employed in the Autonomous City of Buenos Aires's public transport services and combined with data concerning economic and weather conditions, predictive power and most influential features in public transport mobility were measured in two different ways. On one hand, we have driven a traditional SARIMAX time-series model. On the other hand, we have focused on supervised machine learning methods which are designed to enhance prediction capabilities. Given the predominance of the bus as the main public transport service (80% of all trips are made by bus), we particularly focused on this transport mode.

It turns out that the initial suspicion can be confirmed that supervised machine learning algorithms consistently outperformed linear models in predictive power. In terms of most influential features, while machine learning algorithms are often associated with "black-box" results, we found that both type of models show very similar outcomes: As expected, national days, strikes and seasonal effects had a notable impact. However, contrary to previous studies, the only significant variable related to weather conditions was the amount of precipitation.

In terms of elasticity, we evaluated five different demand-elasticities usually discussed in the empirical literature. These are elasticities with respect to the price of petrol, automotive fleet, income, other public transport services and fares.

When it comes to income elasticity, price of petrol and automotive fleet, none of the presented models found a significant relation. However, elasticities with respect to other public transport modes were consistently significant for every model: fare increases in metro and train caused bus passengers to increase.

Finally, we have particularly focused on the own-price elasticity. During the analysis period a persistent inflation affected general prices. In this light, nominal bus fares have had three increases. The last of them was around 80 percent even though the real increase was negligible. This money illusion effect gave us a unique opportunity to evaluate the nominal own-price elasticity. By using a corrected form of the arc elasticity we have compared the SARIMAX model with supervised machine learning algorithms. While none of the supervised models showed a relation between nominal increases and passengers, the SARIMAX formulation subscribes to the empirical rule of thumb of -0.3 (despite being related to real price increases). Moreover, there was an initial shock effect of -0.45, which is consistent with the hypothesis that passengers negatively overreacted to nominal fare increases, but after, they readjusted their consume level.

Chapter 4 examined a typical prediction problem in insurance markets: Fraudulent claims. Specifically, we focused on property claims which have been largely neglected by the literature. By taking advantage of a claim data-set provided by a leading Spanish insurance company, we presented a new methodology to detect fraud. As we discussed in this chapter, the reason for using a semi-supervised algorithm is derived from three key aspects of fraudulent claims: First, data is skewed. It is not surprising that fraud is classified as an abnormal behavior, which means that our data-sets are usually highly unbalanced. Second, as companies have little time to perform exhaustive investigations and they receive thousands of claims per month, there is a substantial mass of claims that are never investigated and, therefore, we do not know which class they belong to. Third, human analysis of fraud cases is poorly adapted to changing patterns.

There are incipient studies which use hybrids of supervised/unsupervised models

to predict fraud. However, they rely on subjective boundaries to define fraud and non-fraud or they assume we always have information about normal behavior.

In this chapter, we tried to solve those three combined problems without making any subjective assumptions that can bias the results. In doing so, we introduced the Cluster Score which measures the abnormal homogeneity in cluster constructions. The methodology involves transmuting unsupervised models to supervised models using this metric, which defines the objective boundaries among clusters.

As we mentioned, we applied this methodology to a real problem of fraud detection among property insurance claims. In the end, 479,454 claims were examined. Our analysis suggests, first, that this methodology considerably increased the number of fraudulent claims detected and reduced the proportion of false positives. Second, the results were not affected by time dynamics (instead, results improved). The real added value, however, is not the ability to capture previously detected cases by the investigation office but rather unsuspicious cases that we have predicted as fraudulent. From a random subsample of 367 claims that were originally classified as unsuspicious (and that we predicted as fraudulent), the investigation office concluded that in fact 333 presented a very high probability to be fraud. This means, in short, that with the methodology proposed we managed to increase fraud detection by 122.8 percent.

Chapter 5 also focused on a well-known insurance problem during the underwriting process, that is, situations in which companies know next to nothing about the risk of their potential new customers. Basic insurance theory suggests that risky customers will not reveal their true nature and, therefore, a suboptimal Pareto equilibrium with an average premium will be reached if no additional incentives are imposed. However, if we assume that not all risky potential customers misreport as insurance theory suggests, we could combine self-reported data with the traditional risk categorization mechanism to solve its inefficiencies when predicting risky potential customers.

In order to shed light in the question if all “bad risks" are misreporters, we used past performance shared data from representative insurers. Thanks to two rich and detailed data sources provided by a Spanish insurance company leader, we had a unique opportunity to evaluate and to validate our main results. First, internal data about customers that signed a vehicle policy permitted us to create proxy variables (by clustering) for the unobservable risk behavior of potential customers. Second, a sample of vehicle insurance policies that were offered to new customers (that may have been transformed into a policy or not) which included details of the offered policy terms. Additionally, it contained various questions related to previous performance which outcome was contrasted with the third source of data: Before subscribing to the policy, the company accessed an external database in which in-

surers shared information regarding previous performance. If a potential customer had not revealed the truth about his or her true nature, an adjusted price was applied. The problem of misreporting is thus reduced to detect who had lied during the underwriting process.

For the empirical exercise, we decided to use a deep variation autoencoder model (VAE) for two reasons: Firstly, insurance companies have considerable amount of rich and reliable data. VAE models obtain compressed representation of the data and, therefore, they can remove undesired features and noise. Secondly, as we demonstrated in this chapter, only a 5.5 percent of potential new customers did not tell the truth. VAE has the advantage that it can be transformed into a powerful semi-supervised outlier detection model. By adopting this methodology, and by combining self-reported data and observable characteristics data,we were able to predict ex-ante between 80 and 87 percent of the risky customers. However, none of the algorithms presented was able to split between risky and non-risky individuals when self-reported data was not used.

In addition, a detailed feature importance analysis showed that the most relevant aspects of policyholders' risk were not related with self-reported prior claims but rather to self-reported years as insured. Our hypothesis was that riskier customers had no incentive to lie about years as insured, because they did not associate it with the final price. However, this variable seems to be implicitly correlated with past behavior. On one hand, a small number of years insured in the last company could reflect a company that choose not to renew the contract to a customer with a bad record. On the other hand, several years as insured could be associated with a good record customer.

We found evidence suggesting that cluster constructed variables related to the customers' zip code and customer characteristics were significant as well. Similarly, the following were also found to be systematically important variables: if the insured was the owner and first driver in the policy, if the customer's age was higher than 65, if the insured was male or female and the number of license years.

The conclusions of the different empirical exercises and the resulting implications of this thesis not only provide reliable results to applications where prediction is more suitable than causal inference but as well a contribution to rethink the way in which we can evaluate traditional economic problems.

6.1 Future Work

There remain some limitations, adaptations and experiments that we seek to address in the future.

In chapter 2, firstly, and as pointed out by Fabra and Fabra Utray (2012), there are two ways in which electric companies could get better prices: (i) by taking off their supply offers during the auction (and, therefore, reducing the competitive pressure), and (ii) by affecting parallel market expectations. We particularly focused on (ii) as micro-data about CESUR auctions was incomplete or unavailable. Counting on this data would let us understand the behavior at the firm level, particularly behaviors related to the mechanisms in which they may have affected prices by retiring quantity offers.

Secondly, our analysis of the economic impact on prices due to collusive patterns is restricted to a control market whose data has only been available since 2013. Therefore, we could only evaluate the last three CESUR auctions. If we are able to get data from the start of the CESUR auctions (2007) or from a similar competitive market as the NordPool, we would get more accurate and robust estimations.

In chapter 3, we have compared several supervised learning, time series and linear models. It would be interesting to apply other increasing popular techniques in the smart card data literature such as support vector machine and deep learning algorithms which could improve our results. It would also be interesting to explore new data about increasing fares and how long-inflation periods have affected the monetary illusion effect.

In chapter 4, and as we previously mentioned, we got very impressing results in terms of predictive power. However, due to the complexity and the nature of the problem, we could not explore statistical models. Despite that, we think that we can still consider less sophisticated approaches to understand what the main motivations of fraud are. As with our methodology we were able to label the non-fraud cases and we have increased the amount of detected fraud, we would like to explore reduced forms of the problem and run several experiments that may reveal the main features of fraud.

Bibliography

Abbring, J. H., Chiappori, P. A. & Pinquet, J. (2003), 'Moral hazard and dynamic insurance data', *Journal of the European Economic Association* **1**(4), 767–820.

Abbring, J. H., Heckman, J. J., Chiappori, P. A. & Pinquet, J. (2003), 'Adverse selection and moral hazard in insurance: Can dynamic data help to distinguish?', *Journal of the European Economic Association* **1**(2-3), 512–521.

Adib, P. & Zarnikau, J. (2006), 'Texas: the most robust competitive market in North America', *Electricity market reform: An international perspective* pp. 383–418.

Afkhami, M., Cormack, L. & Ghoddusi, H. (2017), 'Google search keywords that best predict energy price volatility', *Energy Economics* **67**, 17–27.

Agard, B., Morency, C. & Trépanier, M. (2006), 'Mining public transport user behaviour from smart card data', *IFAC Proceedings Volumes* **39**(3), 399–404.

Aggarwal, C. C. (2015), Outlier analysis, *in* 'Data mining', Springer, pp. 237–263.

Aggarwal, S. K., Saini, L. M. & Kumar, A. (2009), 'Electricity price forecasting in deregulated markets: A review and evaluation', *International Journal of Electrical Power & Energy Systems* **31**(1), 13–22.

Agosti, L., Padilla, A. J. & Requejo, A. (2007), 'El mercado de generación eléctrica en España: estructura, funcionamiento y resultados', *Economía industrial* **364**, 21–37.

Ahmed, M. S. & Cook, A. R. (1979), *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*, number 722.

Ahmed, N. K., Atiya, A. F., Gayar, N. E. & El-Shishiny, H. (2010), 'An empirical comparison of machine learning models for time series forecasting', *Econometric Reviews* **29**(5-6), 594–621.

Ahuja, M. S. & Singh, L. (2017), 'Online fraud detection: A review', *International Research Journal of Engineering and Technology* **4**(7), 2509–2515.

Akerlof, G. A. (1970), 'The market for lemons: Quality uncertainty and the market mechanism', *The Quarterly Journal of Economics* **84**(3), 488–500.

Albalate, D. & Bel, G. (2010), 'What shapes local public transportation in Europe? economics, mobility, institutions, and geography', *Transportation Research Part E: Logistics and Transportation Review* **46**(5), 775–790.

Aleskerov, E., Freisleben, B. & Rao, B. (1997), Cardwatch: A neural network based database mining system for credit card fraud detection, *in* 'Computational Intelligence for Financial Engineering (CIFEr)', IEEE, pp. 220–226.

Ali, A., Kim, J. & Lee, S. (2016), 'Travel behavior analysis using smart card data', *KSCE Journal of Civil Engineering* **20**(4), 1532–1539.

Allaz, B. & Vila, J.-L. (1993), 'Cournot competition, forward markets and efficiency', *Journal of Economic theory* **59**(1), 1–16.

Alsger, A., Assemi, B., Mesbah, M. & Ferreira, L. (2016), 'Validating and improving public transport origin–destination estimation algorithm using smart card fare data', *Transportation Research Part C: Emerging Technologies* **68**, 490–506.

An, J. & Cho, S. (2015), 'Variational autoencoder based anomaly detection using reconstruction probability', *Special Lecture on IE* **2**, 1–18.

Andrews, B. H., Dean, M. D., Swain, R. & Cole, C. (2013), 'Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors', *Society of Actuaries* pp. 1–54.

Andrews, J. T. A., Morton, E. J. & Griffin, L. D. (2016), 'Detecting anomalous data using Auto-Encoders', *International Journal of Machine Learning and Computing* **6**(1), 21–26.

Angrist, J. D. & Krueger, A. B. (1992), 'The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples', *Journal of the American statistical Association* **87**(418), 328–336.

Angrist, J. D. & Krueger, A. B. (2001), 'Instrumental variables and the search for identification: From supply and demand to natural experiments', *Journal of Economic perspectives* **15**(4), 69–85.

Aoyagi, M. (2003), 'Bid rotation and collusion in repeated auctions', *Journal of economic Theory* **112**(1), 79–105.

Arana, P., Cabezudo, S. & Peñalba, M. (2014), 'Influence of weather conditions on transit ridership: A statistical study using data from smartcards', *Transportation research part A: policy and practice* **59**, 1–12.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. (2013), 'An extensive comparative study of cluster validity indices', *Pattern Recognition* **46**(1), 243–256.

Arrow, K. J. (1963), 'Uncertainty and the welfare economics of medical care', *American Economic Review* **53**(5), 941–973.

Arthur, D. & Vassilvitskii, S. (2007), k-means++: The advantages of careful seeding, *in* 'Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms', pp. 1027–1035.

Artis, M., Ayuso, M. & Guillen, M. (1999), 'Modelling different types of automobile insurance fraud behaviour in the Spanish market', *Insurance: Mathematics and Economics* **24**(1-2), 67–81.

Artis, M., Ayuso, M. & Guillen, M. (2002), 'Detection of automobile insurance fraud with discrete choice models and misclassified claims', *Journal of Risk and Insurance* **69**(3), 325–340.

Ascarza, E. (2018), 'Retention futility: Targeting high-risk customers might be ineffective', *Journal of Marketing Research* **55**(1), 80–98.

Atasoy, B., Glerum, A., Hurtubia, R. & Bierlaire, M. (2010), Demand for public transport services: Integrating qualitative and quantitative methods, *in* '10th Swiss Transport Research Conference', number EPFL-CONF-152347.

Athey, S. (2017), 'Beyond prediction: Using big data for policy problems', *Science* **355**(6324), 483–485.

Athey, S. (2018), The impact of machine learning on economics, *in* 'The Economics of Artificial Intelligence: An Agenda', University of Chicago Press.

Athey, S., Bagwell, K. & Sanchirico, C. (2004), 'Collusion and price rigidity', *The Review of Economic Studies* **71**(2), 317–349.

Athey, S. & Imbens, G. W. (2017), 'The state of applied econometrics: Causality and policy evaluation', *Journal of Economic Perspectives* **31**(2), 3–32.

Bagchi, M. & White, P. R. (2005), 'The potential of public transport smart card data', *Transport Policy* **12**(5), 464–474.

Bajari, P. & Ye, L. (2003), 'Deciding between competition and collusion', *Review of Economics and statistics* **85**(4), 971–989.

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M. & White, P. (2004), 'The demand for public transport: a practical guide', *Transportation Research Laboratory Report* **593**.

Banerjee, A. & Dave, R. N. (2004), Validating clusters using the Hopkins statistic, *in* '2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542)', Vol. 1, IEEE, pp. 149–153.

Barnett, V. & Lewis, T. (1994), *Outliers in statistical data*, Wiley Chichester.

Bejani, M. M. & Ghatee, M. (2018), 'A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data', *Transportation Research Part C: Emerging Technologies* **89**, 303–320.

Belhadji, E. B., Dionne, G. & Tarkhani, F. (2000), 'A model for the detection of insurance fraud', *The Geneva Papers on Risk and Insurance-Issues and Practice* **25**(4), 517–538.

Ben-Akiva, M. E. (1973), Structure of passenger travel demand models, PhD thesis, Massachusetts Institute of Technology.

Benjamin, R. M. (2011), Tacit collusion in real-time US electricity auctions, Technical report, USAEE Working Paper.

Bentley, P. J. (2000), Evolutionary, my dear Watson investigating committee-based evolution of fuzzy rules for the detection of suspicious insurance claims, *in* 'Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation', Morgan Kaufmann Publishers Inc., pp. 702–709.

Berck, P. & Villas-Boas, S. B. (2016), 'A note on the triple difference in economic models', *Applied Economics Letters* **23**(4), 239–242.

Berk, R. (2012), *Criminal justice forecasts of risk: A machine learning approach*, Springer Science & Business Media.

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly journal of economics* **119**(1), 249–275.

Björkegren, D. & Grissen, D. (2018), 'Behavior revealed in mobile phone usage predicts loan repayment', *Available at SSRN 2611775* .

Blanco, O. A. (2011), '¿Es competitivo el mercado eléctrico español?: indicadores de abuso de poder de mercado y aplicación al caso de España', *Estudios de economía aplicada* **29**(2), 11–27.

Blumenstock, J., Cadamuro, G. & On, R. (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science* **350**(6264), 1073–1076.

Blythe, P. T. (2004), Improving public transport ticketing through smart cards, *in* 'Proceedings of the Institution of Civil Engineers-Municipal Engineer', Vol. 157, Citeseer, pp. 47–54.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of econometrics* **31**(3), 307–327.

Bollinger, C. R. & David, M. H. (1997), 'Modeling discrete choice with response error: Food stamp participation', *Journal of the American Statistical Association* **92**(439), 827–835.

Bonnel, P. & Chausse, A. (2000), 'Urban travel: Competition and pricing', *Transport reviews* **20**(4), 385–401.

Borenstein, S. & Shepard, A. (1996), Dynamic pricing in retail gasoline markets, Technical report, National Bureau of Economic Research.

Bortoluzzo, A. B., Claro, D. P., Caetano, M. A. L. & Artes, R. (2011), 'Estimating total claim size in the auto insurance industry: a comparison between tweedie and zero-adjusted inverse gaussian distribution', *BAR-Brazilian Administration Review* **8**(1), 37–47.

Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015), *Time series analysis: forecasting and control*, John Wiley & Sons.

Braver, E. R. & Trempel, R. (2004), 'Are older drivers actually at higher risk of involvement in collisions resulting in deaths or non-fatal injuries among their passengers and other road users?', *Injury prevention* **10**(1), 27–32.

Breiman, L. (1996), 'Bagging predictors', *Machine learning* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L. (2017), *Classification and regression trees*, Routledge.

Bresson, G., Dargay, J., Madre, J. L. & Pirotte, A. (2003), 'The main determinants of the demand for public transport: a comparative analysis of England and France using shrinkage estimators', *Transportation Research part A: policy and practice* **37**(7), 605–627.

Bresson, G., Dargay, J., Madre, J. L. & Pirotte, A. (2004), 'Economic and structural determinants of the demand for public transport: an analysis on a panel of French urban areas using shrinkage estimators', *Transportation Research Part A: Policy and Practice* **38**(4), 269–285.

Briand, A.-S., Côme, E., Trépanier, M. & Oukhellou, L. (2017), 'Analyzing year-to-year changes in public transport passenger behaviour using smart card data', *Transportation Research Part C: Emerging Technologies* **79**, 274–289.

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A. & Alpert, M. (2002), 'Fraud classification using principal component analysis of RIDITs', *Journal of Risk and Insurance* **69**(3), 341–371.

Brockett, P. L., Xia, X. & Derrig, R. A. (1998), 'Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud', *Journal of Risk and Insurance* **65**(2), 245–274.

Brown, R. L., Charters, D., Gunz, S. & Haddow, N. (2007), 'Colliding interests–age as an automobile insurance rating variable: Equitable rate-making or unfair discrimination?', *Journal of business ethics* **72**(2), 103–114.

Brugiavini, A. (1993), 'Uncertainty resolution and the timing of annuity purchases', *Journal of Public Economics* **50**(1), 31–62.

Caliński, T. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics-theory and Methods* **3**(1), 1–27.

Card, D. & Krueger, A. (1994), 'A case study of the fast-food industry in New Jersey and Pennsylvania', *American Economic Review* **84**(4), 773–93.

Cardon, J. H. & Hendel, I. (2001), 'Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey', *RAND Journal of Economics* **32**(3), 408–428.

Cartea, A. & Villaplana, P. (2014), An analysis of the main determinants of electricity forward prices and forward risk premia, *in* 'Quantitative Energy Finance', Springer, pp. 215–236.

Catalão, J. P. d. S., Mariano, S. J. P. S., Mendes, V. & Ferreira, L. (2007), 'Short-term electricity prices forecasting in a competitive market: A neural network approach', *Electric Power Systems Research* **77**(10), 1297–1304.

Cawley, J. & Philipson, T. (1999), 'An empirical examination of information barriers to trade in insurance', *American Economic Review* **89**(4), 827–846.

Chandler, D., Levitt, S. D. & List, J. A. (2011), 'Predicting and preventing shootings among at-risk youth', *American Economic Review* **101**(3), 288–92.

Chassang, S. & Ortner, J. (2019), 'Collusion in auctions with constrained bids: Theory and evidence from public procurement', *Journal of Political Economy* **127**(5), 2269–2300.

Che, J. & Wang, J. (2010), 'Short-term electricity prices forecasting based on support vector regression and auto-regressive integrated moving average modeling', *Energy Conversion and Management* **51**(10), 1911–1917.

Che, Y. K. & Kim, J. (2009), 'Optimal collusion-proof auctions', *Journal of Economic Theory* **144**(2), 565–603.

Chen, K., Jiang, J., Zheng, F. & Chen, K. (2018), 'A novel data-driven approach for residential electricity consumption prediction based on ensemble learning', *Energy* **150**, 49–60.

Cherkassky, V. & Ma, Y. (2002), 'Selecting of the loss function for robust linear regression', *Neural computation, NECO* .

Chiappori, P. (1994), Théorie des contrats et économétrie de l'assurance: quelques pistes de recherche, Technical report, Laval-Laboratoire Econometrie.

Chiappori, P. A., Jullien, B., Salanié, B. & Salanie, F. (2006), 'Asymmetric information in insurance: General testable implications', *The RAND Journal of Economics* **37**(4), 783–798.

Chiappori, P. A. & Salanie, B. (2000), 'Testing for asymmetric information in insurance markets', *Journal of Political Economy* **108**(1), 56–78.

Chiappori, P., Jullien, B., Salanie, B. & Salanie, F. (2002), Asymetric information in insurance: some testable implications, Technical report, Working paper.

Cho, M., Hwang, J. & Chen, C. (1995), Customer short term load forecasting by using ARIMA transfer function model, *in* 'Proceedings 1995 International Conference on Energy Management and Power Delivery EMPD'95', Vol. 1, IEEE, pp. 317–322.

Chu, K. K. A. & Chapleau, R. (2008), 'Enriching archived smart card transaction data for transit demand modeling', *Transportation research record* **2063**(1), 63–72.

Ciregan, D., Meier, U. & Schmidhuber, J. (2012), Multi-column deep neural networks for image classification, *in* '2012 IEEE conference on computer vision and pattern recognition', IEEE, pp. 3642–3649.

Cohen, A. (2005), 'Asymmetric information and learning: Evidence from the automobile insurance market', *Review of Economics and statistics* **87**(2), 197–207.

Cohen, A. (2008), Asymmetric learning in repeated contracting: An empirical study, Technical report, National Bureau of Economic Research.

Cohen, A. & Siegelman, P. (2010), 'Testing for adverse selection in insurance markets', *Journal of Risk and insurance* **77**(1), 39–84.

Comisión Nacional de los Mercados y la Competencia (2014), 'Informe sobre el desarrollo de la 25ª subasta CESUR previsto en el artículo 14.3 de la orden ITC/1659/2009, de 22 de Junio'.

Conejo, A. J., Contreras, J., Espinola, R. & Plazas, M. A. (2005), 'Forecasting electricity prices for a day-ahead pool-based electric energy market', *International journal of forecasting* **21**(3), 435–462.

Conejo, A. J., Plazas, M. A., Espinola, R. & Molina, A. B. (2005), 'Day-ahead electricity price forecasting using the wavelet transform and ARIMA models', *IEEE transactions on power systems* **20**(2), 1035–1042.

Contreras, J., Espinola, R., Nogales, F. J. & Conejo, A. J. (2003), 'ARIMA models to predict next-day electricity prices', *IEEE transactions on power systems* **18**(3), 1014–1020.

Cooper, R. & Hayes, B. (1987), 'Multi-period insurance contracts', *International Journal of Industrial Organization* **5**(2), 211–231.

Cox, E. (1995), 'A fuzzy system for detecting anomalous behaviors in healthcare provider claims', *Intelligent Systems for Finance and Business* pp. 111–134.

Cozzolino, D. & Verdoliva, L. (2016), Single-image splicing localization through autoencoder-based anomaly detection, *in* '2016 IEEE International Workshop on Information Forensics and Security (WIFS)', pp. 1–6.

Crocker, K. J. & Snow, A. (1985), 'The efficiency of competitive equilibria in insurance markets with asymmetric information', *Journal of Public economics* **26**(2), 207–219.

Crocker, K. J. & Snow, A. (1986), 'The efficiency effects of categorical discrimination in the insurance industry', *Journal of Political Economy* **94**(2), 321–344.

Crocker, K. J. & Snow, A. (2013), The theory of risk classification, *in* 'Handbook of insurance', Springer, pp. 281–313.

Crôtte, A. (2008), Estimation of transport related demand elasticities in Mexico city: An application to road user charging, PhD thesis, Centre for Transport Studies.

Cuaresma, J. C., Hlouskova, J., Kossmeier, S. & Obersteiner, M. (2004), 'Forecasting electricity spot-prices using linear univariate time-series models', *Applied Energy* **77**(1), 87–106.

Dahlby, B. (1992), Testing for asymmetric information in Canadian automobile insurance, *in* 'Contributions to Insurance Economics', Springer, pp. 423–443.

Dahlby, B. G. (1983), 'Adverse selection and statistical discrimination: An analysis of Canadian automobile insurance', *Journal of Public Economics* **20**(1), 121–130.

D'Arcy, S. P. & Doherty, N. A. (1990), 'Adverse selection, private information, and lowballing in insurance markets', *Journal of Business* **63**(2), 145–164.

Dau, H. & Song, A. (2014), 'Anomaly detection using replicator neural networks trained on examples of one class', *Proceedings of 10th International Conference on Simulated Evolution And Learning* pp. 311–322.

David, M. (2015), Automobile insurance pricing with generalized linear models, *in* 'Proceedings in GV-Global Virtual Conference', number 1.

Davidoff, T. & Welke, G. (2004), Selection and moral hazard in the reverse mortgage market, Technical report.

Davies, D. L. & Bouldin, D. W. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2), 224–227.

D'Avino, D., Cozzolino, D., Poggi, G. & Verdoliva, L. (2017), 'Autoencoder with recurrent neural networks for video forgery detection', *Electronic Imaging* (7), 92–99.

De Frutos, M. A. & Fabra, N. (2011), 'Endogenous capacities and price competition: The role of demand uncertainty', *International Journal of Industrial Organization* **29**(4), 399–411.

de Grange, L., González, F., Muñoz, J. C. & Troncoso, R. (2013), 'Aggregate estimation of the price elasticity of demand for public transport in integrated fare systems: The case of Transantiago', *Transport Policy* **29**, 178–185.

de Wit, G. W. (1982), 'Underwriting and uncertainty', *Insurance: Mathematics and Economics* **1**(4), 277–285.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K., Le, Q. V. et al. (2012), Large scale distributed deep networks, *in* 'Advances in neural information processing systems', pp. 1223–1231.

Defeuilley, C. (2009), 'Retail competition in electricity markets', *Energy Policy* **37**(2), 377–386.

Dickey, D. A. & Fuller, W. A. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American statistical association* **74**(366a), 427–431.

Dionne, G. (1992), Adverse selection and repeated insurance contracts, *in* 'Foundations of Insurance Economics', Springer, pp. 407–423.

Dionne, G. & Doherty, N. A. (1994), 'Adverse selection, commitment, and renegotiation: Extension to and evidence from insurance markets', *Journal of political Economy* **102**(2), 209–235.

Dionne, G., Doherty, N. & Nathalie, F. (2000), *Adverse Selection in Insurance Markets*, Springer.

Dionne, G., Gouriéroux, C. & Vanasse, C. (1999), Evidence of adverse selection in automobile insurance markets, *in* 'Automobile Insurance: Road safety, new drivers, risks, insurance fraud and regulation', Springer, pp. 13–46.

Dionne, G. & Lasserre, P. (1985), 'Adverse selection, repeated insurance contracts and announcement strategy', *The Review of Economic Studies* **52**(4), 719–723.

Dionne, G. & Lasserre, P. (1987), 'Adverse selection and finite-horizon insurance contracts', *European Economic Review* **31**(4), 843–861.

Dionne, G., Michaud, P. C. & Dahchour, M. (2013), 'Separating moral hazard from adverse selection and learning in automobile insurance: longitudinal evidence from France', *Journal of the European Economic Association* **11**(4), 897–917.

Dionne, G. & Vanasse, C. (1992), 'Automobile insurance ratemaking in the presence of asymmetrical information', *Journal of Applied Econometrics* **7**(2), 149–165.

Doerpinghaus, H. I., Schmit, J. T. & Yeh, J. J.-H. (2008), 'Age and gender effects on auto liability insurance payouts', *Journal of Risk and Insurance* **75**(3), 527–550.

Dou, M., He, T., Yin, H., Zhou, X., Chen, Z. & Luo, B. (2015), Predicting passengers in public transportation using smart card data, *in* 'Australasian Database Conference', Springer, pp. 28–40.

Duchi, J., Hazan, E. & Singer, Y. (2011), 'Adaptive subgradient methods for online learning and stochastic optimization', *Journal of Machine Learning Research* **12**(Jul), 2121–2159.

Durbin, J. & Watson, G. S. (1950), 'Testing for serial correlation in least squares regression: I', *Biometrika* **37**(3/4), 409–428.

Durbin, J. & Watson, G. S. (1951), 'Testing for serial correlation in least squares regression. II', *Biometrika* **38**(1/2), 159–177.

Durbin, J. & Watson, G. S. (1971), 'Testing for serial correlation in least squares regression. III', *Biometrika* **58**(1), 1–19.

E. Rumelhart, D., E. Hinton, G. & J. Williams, R. (1986), *Learning Internal Representation by Error Propagation*, Vol. Vol. 1.

Efthymiou, D. & Antoniou, C. (2017), 'Understanding the effects of economic crisis on public transport users' satisfaction and demand', *Transport Policy* **53**, 89–97.

Einav, L., Finkelstein, A. & Levin, J. (2010), 'Beyond testing: Empirical models of insurance markets', *Annual Review of Economics* **2**(1), 311–336.

Ellison, G. (1994), 'Theories of cartel stability and the joint executive committee', *The Rand journal of economics* **25**(1), 37–57.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation', *Econometrica: Journal of the Econometric Society* **50**(4), 987–1007.

Engstrom, R., Hersh, J. & Newhouse, D. (2016), Poverty in HD: What does high resolution satellite imagery reveal about economic welfare, Technical report, Working paper.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *in* 'Kdd', Vol. 96, pp. 226–231.

Europe, Insurance (2013), 'The impact of insurance fraud', *Brussels: Insurance Europe* .

Fabra, N. (2003), 'Tacit collusion in repeated auctions: uniform versus discriminatory', *The Journal of Industrial Economics* **51**(3), 271–293.

Fabra, N. (2006), 'Collusion with capacity constraints over the business cycle', *International Journal of Industrial Organization* **24**(1), 69–81.

Fabra, N. & García, A. (2015), 'Dynamic price competition with switching costs', *Dynamic Games and Applications* **5**(4), 540–567.

Fabra, N. & Reguant, M. (2014), 'Pass-through of emissions costs in electricity markets', *American Economic Review* **104**(9), 2872–99.

Fabra, N. & Toro, J. (2005), 'Price wars and collusion in the Spanish electricity market', *International Journal of Industrial Organization* **23**(3-4), 155–181.

Fabra, N. & Utray, J. F. (2010), 'Competencia y poder de mercado en los mercados eléctricos', *Cuadernos Económicos de ICE* (79).

Fabra, N. & Utray, J. F. (2012), 'El déficit tarifario en el sector eléctrico español', *Papeles de economía española* (134), 88–100.

Fabra, N., Von der Fehr, N.-H. & Harbord, D. (2002), 'Modeling electricity auctions', *The Electricity Journal* **15**(7), 72–81.

Fabra, N., von der Fehr, N.-H. & Harbord, D. (2006), 'Designing electricity auctions', *The RAND Journal of Economics* **37**(1), 23–46.

Fan, S., Mao, C. & Chen, L. (2007), 'Next-day electricity-price forecasting using a hybrid network', *IET generation, Transmission & Distribution* **1**(1), 176–182.

Fezzi, C. (2007), Econometric models for the analysis of electricity markets, PhD thesis, ALMA.

Figueiredo, M. A. T. & Jain, A. K. (2002), 'Unsupervised learning of finite mixture models', *Transactions on Pattern Analysis and Machine Intelligence* **24**(3), 381–396.

Finkelstein, A. & McGarry, K. (2006), 'Multiple dimensions of private information: evidence from the long-term care insurance market', *American Economic Review* **96**(4), 938–958.

Finkelstein, A., McGarry, K. & Sufi, A. (2005), 'Dynamic inefficiencies in insurance markets: Evidence from long-term care insurance', *American Economic Review* **95**(2), 224–228.

Finkelstein, A. & Poterba, J. (2002), 'Selection effects in the United Kingdom individual annuities market', *The Economic Journal* **112**(476), 28–50.

Finkelstein, A. & Poterba, J. (2004), 'Adverse selection in insurance markets: Policyholder evidence from the UK annuity market', *Journal of Political Economy* **112**(1), 183–208.

Finkelstein, A. & Poterba, J. (2006), Testing for asymmetric information using 'unused observables' in insurance markets: Evidence from the UK annuity market, NBER Working Papers 12112, National Bureau of Economic Research, Inc.

FitzRoy, F. & Smith, I. (1998), 'Public transport demand in Freiburg: why did patronage double in a decade?', *Transport policy* **5**(3), 163–173.

Fombaron, N. (1997), No-commitment and dynamic contracts in competitive insurance markets with adverse selection, Technical report, THEMA, Université de Cergy-Pontoise.

Fombaron, N. (2000), Renegotiation-proof contracts in insurance markets with asymmetric information, Technical report, Working Paper, Thema, Université de Cergy-Pontoise.

Foster, D. P. & Stine, R. A. (2004), 'Variable selection in data mining: Building a predictive model for bankruptcy', *Journal of the American Statistical Association* **99**(466), 303–313.

Freifelder, L. R. (1985), 'Measuring the impact of merit rating on ratemaking efficiency', *Journal of Risk and Insurance* **52**(4), 607–626.

Freund, Y., Schapire, R. E. et al. (1996), Experiments with a new boosting algorithm, *in* 'ICML', Vol. 96, Citeseer, pp. 148–156.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* **29**(5), 1189–1232.

Fuller, W. A. (1976), *Introduction to statistical time series*, Vol. 428, John Wiley & Sons.

Furió, D. & Meneu, V. (2010), 'Expectations and forward risk premium in the Spanish deregulated power market', *Energy Policy* **38**(2), 784–793.

García-Ferrer, A., Bujosa, M., de Juan, A. & Poncela, P. (2006), 'Demand forecast and elasticities estimation of public transport', *Journal of Transport Economics and Policy (JTEP)* **40**(1), 45–67.

Gepp, A., Wilson, J., Kumar, K. & Bhattacharya, S. (2012), 'A comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection', *Journal of data science: JDS* **10**, 537–561.

Geurts, P., Ernst, D. & Wehenkel, L. (2006*a*), 'Extremely randomized trees', *Machine learning* **63**(1), 3–42.

Geurts, P., Ernst, D. & Wehenkel, L. (2006*b*), 'Extremely randomized trees', *Machine Learning* **63**(1), 3–42.

Ghoddusi, H., Creamer, G. G. & Rafizadeh, N. (2019), 'Machine learning in energy economics and finance: A review', *Energy Economics* **81**.

Glaeser, E. L., Hillis, A., Kominers, S. D. & Luca, M. (2016), 'Crowdsourcing city government: Using tournaments to improve inspection accuracy', *American Economic Review* **106**(5), 114–18.

Goel, S., Rao, J. M., Shroff, R. et al. (2016), 'Precinct or prejudice? understanding racial disparities in New York City's stop-and-frisk policy', *The Annals of Applied Statistics* **10**(1), 365–394.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S. & Mohammadian, A. (2018), 'Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model', *Travel Behaviour and Society* **10**, 21–32.

Gong, M., Fei, X., Wang, Z. H. & Qiu, Y. J. (2014), 'Sequential framework for short-term passenger flow prediction at bus stop', *Transportation Research Record* **2417**(1), 58–66.

Goodwin, P. B. (1992), 'A review of new demand elasticities with special reference to short and long run effects of price changes', *Journal of transport economics and policy* **26**(2), 155–169.

Goto, M. & Karolyi, G. (2004), Understanding electricity price volatility within and across markets, Technical report, Ohio State University.

Graham, D. A. & Marshall, R. C. (1987), 'Collusive bidder behavior at single-object second-price and English auctions', *Journal of Political economy* **95**(6), 1217–1239.

Graham, D. J., Crotte, A. & Anderson, R. J. (2009), 'A dynamic panel analysis of urban metro demand', *Transportation Research Part E: Logistics and Transportation Review* **45**(5), 787–794.

Green, E. J. & Porter, R. H. (1984), 'Noncooperative collusion under imperfect price information', *Econometrica: Journal of the Econometric Society* **52**(1), 87–100.

Green, R. (1999), 'The electricity contract market in England and Wales', *The Journal of Industrial Economics* **47**(1), 107–124.

Green, R. J. (2004), 'Retail competition and electricity contracts'.

Green, R. J. & Newbery, D. M. (1992), 'Competition in the British electricity spot market', *Journal of political economy* **100**(5), 929–953.

Grimmer, J. & Stewart, B. M. (2013), 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political analysis* **21**(3), 267–297.

Gruber, J. (1994), 'The incidence of mandated maternity benefits', *The American economic review* **84**(3), 622–641.

Gu, Z., Saberi, M., Sarvi, M. & Liu, Z. (2018), 'A big data approach for clustering and calibration of link fundamental diagrams for large-scale network simulation applications', *Transportation Research Part C: Emerging Technologies* **94**, 151–171.

Guelman, L. (2012), 'Gradient boosting trees for auto insurance loss cost modeling and prediction', *Expert Systems with Applications* **39**(3), 3659–3667.

Guo, F., Polak, J. W. & Krishnan, R. (2018), 'Predictor fusion for short-term traffic forecasting', *Transportation Research Part C: Emerging Technologies* **92**, 90–100.

Hagenauer, J. & Helbich, M. (2017), 'A comparative study of machine learning classifiers for modeling travel mode choice', *Expert Systems with Applications* **78**, 273–282.

Haltiwanger, J. & Harrington Jr, J. E. (1991), 'The impact of cyclical demand movements on collusive behavior', *The RAND Journal of Economics* **22**(1), 89–106.

Hammad, K., Fakharaldien, M., Zain, J. & Majid, M. (2015), Big data analysis and storage, *in* 'International Conference on Operations Excellence and Service Engineering', pp. 10–11.

Hanly, M., Dargay, J. & Goodwin, P. (2002), 'Review of income and price elasticities in the demand for road traffic', *Department for Transport, London* .

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Hauer, E. (1971), 'Fleet selection for public transportation routes', *Transportation Science* **5**(1), 1–21.

He, D. (2009), 'The life insurance market: Asymmetric information revisited', *Journal of Public Economics* **93**(9-10), 1090–1097.

He, H., Bai, Y., Garcia, E. A. & Li, S. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *in* 'IJCNN (World Congress on Computational Intelligence)', IEEE, pp. 1322–1328.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Hellwig, M. (1986), A sequential approach to modelling competition in markets with adverse selection, *in* 'University of Bonn Discussion paper', Vol. 303.

Hendel, I. & Lizzeri, A. (2003), 'The role of commitment in dynamic contracts: Evidence from life insurance', *The Quarterly Journal of Economics* **118**(1), 299–328.

Hendren, N. (2013), 'Private information and insurance rejections', *Econometrica* **81**(5), 1713–1762.

Hensher, D. A. (1998), 'Establishing a fare elasticity regime for urban passenger transport', *Journal of Transport Economics and Policy* **32**(2), 221–246.

Herraiz, A. C. & Monroy, C. R. (2012), 'Evaluation of the trading development in the Iberian Energy Derivatives Market', *Energy policy* **51**, 973–984.

Hinton, G. E., Krizhevsky, A. & Wang, S. D. (2011), Transforming auto-encoders, *in* 'International Conference on Artificial Neural Networks', Springer, pp. 44–51.

Hinton, G. E. & Salakhutdinov, R. R. (2006), 'Reducing the dimensionality of data with neural networks', *American Association for the Advancement of Science* **313**(5786), 504–507.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hodge, V. & Austin, J. (2004), 'A survey of outlier detection methodologies', *Artificial intelligence review* **22**(2), 85–126.

Hodrick, R. J. & Prescott, E. C. (1997), 'Postwar US business cycles: an empirical investigation', *Journal of Money, credit, and Banking* **29**(1), 1–16.

Holmgren, J. (2007), 'Meta-analysis of public transport demand', *Transportation Research Part A: Policy and Practice* **41**(10), 1021–1035.

Hopkins, B. & Skellam, J. G. (1954), 'A new method for determining the type of distribution of plant individuals', *Annals of Botany* **18**(2), 213–227.

Hosios, A. J. & Peters, M. (1989), 'Repeated insurance contracts with adverse selection and limited commitment', *The Quarterly Journal of Economics* **104**(2), 229–253.

Hoy, M. (1982), 'Categorizing risks in the insurance industry', *The Quarterly Journal of Economics* **97**(2), 321–336.

Huang, S. (1997), 'Short-term load forecasting using threshold autoregressive models', *IEE Proceedings-Generation, Transmission and Distribution* **144**(5), 477–481.

Huang, S. J. & Shih, K. R. (2003), 'Short-term load forecasting via ARMA model identification including non-Gaussian process considerations', *IEEE Transactions on power systems* **18**(2), 673–679.

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T. & Wang, F.-Y. (2018), 'Modeling real-time human mobility based on mobile phone and transportation data fusion', *Transportation Research Part C: Emerging Technologies* **96**, 251–269.

Hyndman, R. J. & Athanasopoulos, G. (2018), *Forecasting: principles and practice*, OTexts.

Ingvardson, J. B., Nielsen, O. A., Raveau, S. & Nielsen, B. F. (2018), 'Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis', *Transportation Research Part C: Emerging Technologies* **90**, 292–306.

Ishii, R. (2008), 'Collusion in repeated procurement auction: a study of a paving market in Japan', *ISER Discussion Paper* **710**.

Israel, M. (2004), Do we drive more safely when accidents are more expensive? identifying moral hazard from experience rating schemes, Technical report, Northwestern University, Center for the Study of Industrial Organization.

Jakovčević, D. & Žaja, M. M. (2014), 'Underwriting risks as determinants of insurance cycles: Case of Croatia', *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* **8**(5), 1251 – 1258.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science* **353**(6301), 790–794.

Joskow, P. L. (2000), 'Why do we need electricity retailers?; or, can you get it cheaper wholesale?', *MIT Center for Energy and Environmental Policy Research* .

Jung, J. & Sohn, K. (2017), 'Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data', *IET Intelligent Transport Systems* **11**(6), 334–339.

Kang, J. S., Kuznetsova, P., Luca, M. & Choi, Y. (2013), Where not to eat? improving public policy by predicting hygiene inspections using online reviews, *in* 'Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing', pp. 1443–1448.

Kašćelan, L., Kašćelan, V. & Jovanović, M. (2015), 'Hybrid support vector machine rule extraction method for discovering the preferences of stock market investors: Evidence from Montenegro', *Intelligent Automation & Soft Computing* **21**(4), 503–522.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017), Lightgbm: A highly efficient gradient boosting decision tree, *in* 'Advances in Neural Information Processing Systems', pp. 3146–3154.

Ke, J., Zheng, H., Yang, H. & Chen, X. M. (2017), 'Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach', *Transportation Research Part C: Emerging Technologies* **85**, 591–608.

Kim, J., Ong, A. & Overill, R. E. (2003), Design of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector, *in* 'Evolutionary Computation', Vol. 1, IEEE, pp. 405–412.

Kingma, D. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *International Conference on Learning Representations* **14**.

Kingma, D. & Welling, M. (2014), Auto-encoding variational bayes.

Klaiber, H. A. & von Haefen, R. H. (2019), 'Do random coefficients and alternative specific constants improve policy analysis? an empirical investigation of model fit and prediction', *Environmental and Resource Economics* **73**, 75–91.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review* **105**(5), 491–95.

Kokkinaki, A. I. (1997), On atypical database transactions: identification of probable frauds using machine learning for user profiling, *in* 'Knowledge and Data Engineering Exchange Workshop, 1997. Proceedings', IEEE, pp. 107–113.

Kraft, G. & Wohl, M. (1967), 'New directions for passenger demand analysis and forecasting', *Transportation Research* **1**(3), 205–230.

Kremers, H., Nijkamp, P. & Rietveld, P. (2002), 'A meta-analysis of price elasticities of transport demand in a general equilibrium framework', *Economic Modelling* **19**(3), 463–485.

Krizhevsky, A. & Hinton, G. E. (2011), Using very deep autoencoders for content-based image retrieval, *in* 'ESANN'.

Kumar, P., Khani, A. & He, Q. (2018), 'A robust method for estimating transit passenger trajectories using automated data', *Transportation Research Part C: Emerging Technologies* **95**, 731–747.

Kunreuther, H., Meszaros, J., Hogarth, R. M. & Spranca, M. (1995), 'Ambiguity and underwriter decision processes', *Journal of Economic Behavior & Organization* **26**(3), 337–352.

Kunreuther, H. & Pauly, M. (1985), Market equilibrium with private knowledge, *in* 'Foundations of Insurance Economics', Springer, pp. 424–443.

Kusakabe, T. & Asakura, Y. (2014), 'Behavioural data mining of transit smart card data: A data fusion approach', *Transportation Research Part C: Emerging Technologies* **46**, 179–191.

Lago, J., De Ridder, F. & De Schutter, B. (2018), 'Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms', *Applied Energy* **221**, 386–405.

Last, M., Maimon, O. & Minkov, E. (2002), 'Improving stability of decision trees', *International Journal of Pattern Recognition and Artificial Intelligence* **16**(02), 145–159.

Lee, C. C. & Chiu, Y. B. (2011), 'Electricity demand elasticities and temperature: Evidence from panel smooth transition regression with instrumental variable approach', *Energy Economics* **33**(5), 896–902.

Lei, J. Z. & Ghorbani, A. A. (2012), 'Improved competitive learning neural networks for network intrusion and fraud detection', *Neurocomputing* **75**(1), 135–145.

Lemaire, J. (1985), *Automobile insurance: actuarial models*, Vol. 4, Springer Science & Business Media.

Lemaire, J. (1995), *Bonus-malus systems in automobile insurance*, Vol. 19, Springer science & business media.

Leon, A. & Rubia, A. (2004*a*), *Forecasting time-varying covariance matrices in the intradaily spot market of Argentina, in (ed.: DW Bunn) Modelling prices in competitive electricity markets*, Wiley & Sons, Chichester.

León, A. & Rubia, A. (2004*b*), 'Testing for weekly seasonal unit roots in the Spanish power pool', *Modelling Prices in Competitive Electricity Markets. Wiley Series in Financial Economics* pp. 131–145.

Li, H. & Chen, X. (2016), 'Unifying time reference of smart card data using dynamic time warping', *Procedia engineering* **137**, 513–522.

Li, T., Sun, D., Jing, P. & Yang, K. (2018), 'Smart card data mining of public transport destination: A literature review', *Information* **9**(1), 18.

Lin, W. M., Gow, H. J. & Tsai, M. T. (2010), 'An enhanced radial basis function network for short-term electricity price forecasting', *Applied Energy* **87**(10), 3226–3234.

Litman, T. (2012), Understanding transport demands and elasticities: How prices and other factors affect travel behavior, Technical report, Victoria Transport Policy Institute.

Liu, F. T., Ting, K. M. & Zhou, Z. (2008*a*), Isolation forest, *in* '2008 Eighth IEEE International Conference on Data Mining', pp. 413–422.

Liu, F. T., Ting, K. M. & Zhou, Z. H. (2008*b*), Isolation forest, *in* 'International Conference on Data Mining', IEEE, pp. 413–422.

Liu, L. & Chen, R.-C. (2017), 'A novel passenger flow prediction model using deep learning methods', *Transportation Research Part C: Emerging Technologies* **84**, 74–91.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010), Understanding of internal clustering validation measures, *in* '2010 IEEE International Conference on Data Mining', IEEE, pp. 911–916.

Liu, Y., Liu, Z. & Jia, R. (2019), 'DeepPF: A deep learning based architecture for metro passenger flow prediction', *Transportation Research Part C: Emerging Technologies* **101**, 18–34.

Liu, Y., Wang, B.-J. & Lv, S.-G. (2014), 'Using multi-class AdaBoost tree for prediction frequency of auto insurance', *Journal of Applied Finance and Banking* **4**(5), 45.

Ljung, G. M. & Box, G. E. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297–303.

Lloyd, S. (1982), 'Least squares quantization in PCM', *IEEE transactions on information theory* **28**(2), 129–137.

Loxley, C. & Salant, D. (2004), 'Default service auctions', *Journal of Regulatory Economics* **26**(2), 201–229.

Lyudchik, O. (2016), Outlier detection using autoencoders, Technical report, CERN-STUDENTS-Note-2016-079.

Ma, J., Chan, J., Ristanoski, G., Rajasegarar, S. & Leckie, C. (2019), 'Bus travel time prediction with real-time traffic information', *Transportation Research Part C: Emerging Technologies* **105**, 536–549.

Ma, X., Liu, C., Wen, H., Wang, Y. & Wu, Y.-J. (2017), 'Understanding commuting patterns using transit smart card data', *Journal of Transport Geography* **58**, 135–145.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F. & Liu, J. (2013), 'Mining smart card data for transit riders' travel patterns', *Transportation Research Part C: Emerging Technologies* **36**, 1–12.

Mahan, M. Y., Chorn, C. R. & Georgopoulos, A. P. (2015), White noise test: detecting autocorrelation and nonstationarities in long time series after ARIMA modeling, *in* 'Proceedings 14th Python in Science Conference (Scipy 2015), Austin, TX'.

Major, J. A. & Riedinger, D. R. (1992), 'A hybrid knowledge/statistical-based system for the detection of fraud', *International Journal of Intelligent Systems* **7**(7), 687–703.

Manevitz, L. M. & Yousef, M. (2001), 'One-class SVMs for document classification', *Journal of Machine Learning Research* **2**, 139–154.

Marshall, R. C. & Marx, L. M. (2009), 'The vulnerability of auctions to bidder collusion', *The Quarterly Journal of Economics* **124**(2), 883–910.

Matas, A. (2004), 'Demand and revenue implications of an integrated public transport policy: the case of Madrid', *Transport Reviews* **24**(2), 195–217.

McAfee, R. P. & McMillan, J. (1992), 'Bidding rings', *The American Economic Review* **82**(3), 579–599.

McMillan, J. (1991), 'Dango: Japan's price-fixing conspiracies', *Economics & Politics* **3**(3), 201–218.

Meyer, B. D. (1995), 'Natural and quasi-experiments in economics', *Journal of business & economic statistics* **13**(2), 151–161.

Milenković, M., Švadlenka, L., Melichar, V., Bojović, N. & Avramović, Z. (2018), 'SARIMA modelling approach for railway passenger flow forecasting', *Transport* **33**(5), 1113–1120.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence* **267**, 1–38.

Misiorek, A., Trueck, S. & Weron, R. (2006), 'Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models', *Studies in Nonlinear Dynamics & Econometrics* **10**(3).

Miyazaki, H. (1977), 'The rat race and internal labor markets', *The Bell Journal of Economics* **8**(2), 394–418.

Molnar, C. (2018), 'Interpretable machine learning', *A Guide for Making Black Box Models Explainable* .

Morency, C., Trépanier, M. & Agard, B. (2007), 'Measuring transit use variability with smart-card data', *Transport Policy* **14**(3), 193–203.

Mullainathan, S. & Spiess, J. (2017), 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

Munizaga, M. A. & Palma, C. (2012), 'Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile', *Transportation Research Part C: Emerging Technologies* **24**, 9–18.

Murad, U. & Pinkas, G. (1999), Unsupervised profiling for identifying superimposed fraud, *in* 'European Conference on Principles of Data Mining and Knowledge Discovery', Springer, pp. 251–261.

Murtaugh, C. M., Kemper, P. & Spillman, B. C. (1995), 'Risky business: Long-term care insurance underwriting', *Inquiry* **32**(3), 271–284.

Naik, N., Raskar, R. & Hidalgo, C. A. (2016), 'Cities are physical too: Using computer vision to measure the quality and impact of urban appearance', *American Economic Review* **106**(5), 128–32.

NERA (2007), 'Review of the effectiveness of energy retail market competition in South Australia', *National Economic Reasearch Associates Report* **Phase II and III**.

Neumark, D. & Wascher, W. (1994), 'Employment effects of minimum and subminimum wages: Reply to Card, Katz, and Krueger', *ILR Review* **47**(3), 497–512.

Newbold, P. & Granger, C. W. (1974), 'Experience with forecasting univariate time series and the combination of forecasts', *Journal of the Royal Statistical Society: Series A (General)* **137**(2), 131–146.

Nian, K., Zhang, H., Tayal, A., Coleman, T. & Li, Y. (2016), 'Auto insurance fraud detection using unsupervised spectral ranking for anomaly', *The Journal of Finance and Data Science* **2**(1), 58–75.

Nijkamp, P. & Pepping, G. (1998), 'Meta-analysis for explaining the variance in public transport demand elasticities in europe', *Journal of Transportation and Statistics* **1**(1), 1–14.

Nils-Henrik, M. & Harbord, D. (1993), *Spot market competition in the UK electricity industry*, Vol. 103, Oxford University Press Oxford, UK.

Nilssen, T. (2000), 'Consumer lock-in with asymmetric information', *International Journal of Industrial Organization* **18**(4), 641–666.

Nowicka-Zagrajek, J. & Weron, R. (2002), 'Modeling electricity loads in California: ARMA models with hyperbolic noise', *Signal Processing* **82**(12), 1903–1915.

Nunes, A. A., Dias, T. G. & e Cunha, J. F. (2016), 'Passenger journey destination estimation from automated fare collection system data using spatial validation', *IEEE Transactions on Intelligent Transportation Systems* **17**(1), 133–142.

Ohlsson, E. & Johansson, B. (2010), *Non-life insurance pricing with generalized linear models*, Vol. 2, Springer.

Omrani, H. (2015), 'Predicting travel mode of individuals by machine learning', *Transportation Research Procedia* **10**, 840–849.

Omrani, H., Charif, O., Gerber, P., Awasthi, A. & Trigano, P. (2013), 'Prediction of individual travel mode with evidential neural network model', *Transportation Research Record* **2399**(1), 1–8.

Oort, N., Brands, T. & de Romph, E. (2015), 'Short-term prediction of ridership on public transport with smart card data', *Transportation Research Record: Journal of the Transportation Research Board* **2535**, 105–111.

Oum, T. H., Waters, W. G. & Yong, J.-S. (1992), 'Concepts of price elasticities of transport demand and recent empirical estimates: an interpretative survey', *Journal of Transport Economics and policy* **26**(2), 139–154.

Paisley, J., Blei, D. & Jordan, M. (2012), 'Variational bayesian inference with stochastic search', *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* **2**.

Paula, E. L., Ladeira, M., Carvalho, R. N. & Marzagão, T. (2016), Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering, *in* '2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)', IEEE, pp. 954–960.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J. & White, P. (2006), 'The demand for public transport: The effects of fares, quality of service, income and car ownership', *Transport policy* **13**(4), 295–306.

Pavlov, G. (2008), 'Auction design in the presence of collusion', *Theoretical Economics* **3**(3), 383–429.

Pelletier, M. P., Trépanier, M. & Morency, C. (2011), 'Smart card data use in public transit: A literature review', *Transportation Research Part C: Emerging Technologies* **19**(4), 557–568.

Peña, J. I. & Rodriguez, R. (2018), 'Default supply auctions in electricity markets: Challenges and proposals', *Energy policy* **122**, 142–151.

Perez-Mora, N., Alomar, M. L. & Martinez-Moll, V. (2018), 'Spanish energy market: Overview towards price forecast', *International Journal of Energy Economics and Policy* **8**(3), 1–7.

Philipson, T. & Cawley, J. (1999), 'An empirical examination of information barriers to trade in insurance', *American Economic Review* **89**(4), 827–846.

Phua, C., Alahakoon, D. & Lee, V. (2004), 'Minority report in fraud detection: classification of skewed data', *ACM SIGKDD Explorations Newsletter* **6**(1), 50–59.

Phua, C., Lee, V., Smith, K. & Gayler, R. (2010), 'A comprehensive survey of data mining-based fraud detection research', *ArXiv* **abs/1009.6119**.

Pischke, J. S. (2005), 'Empirical methods in applied economics', *Lecture Notes. London School of Economics* .

Pitombo, C. S., de Souza, A. D. & Lindner, A. (2017), 'Comparing decision tree algorithms to estimate intercity trip distribution', *Transportation Research Part C: Emerging Technologies* **77**, 16–32.

Polat, C. (2012), 'The demand determinants for urban public transport services: a review of the literature', *Journal of Applied Sciences* **12**(12), 1211–1231.

Porter, R. H. (1983), 'A study of cartel stability: the Joint Executive Committee, 1880-1886', *The Bell Journal of Economics* **14**(2), 301–314.

Porter, R. H. & Zona, J. D. (1993), 'Detection of bid rigging in procurement auctions', *Journal of political economy* **101**(3), 518–538.

Porter, R. H. & Zona, J. D. (1999), Ohio school milk markets: An analysis of bidding, Technical report, National Bureau of Economic Research.

Poterba, J. M. & Summers, L. H. (1995), 'Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification', *The Review of Economics and Statistics* **77**(2), 207–216.

Powers, M., Gao, F. & Wang, J. (2009), 'Adverse selection or advantageous selection? Risk and underwriting in China's health-insurance market', **44**, 505–510.

Preston, J. (1998), *Public Transport Elasticities: Time for a Re-think?*, University of Oxford, Transport Studies Unit.

Protopapadakis, E., Voulodimos, A., Doulamis, A., Doulamis, N., Dres, D. & Bimpas, M. (2017), 'Stacked autoencoders for outlier detection in over-the-horizon radar signals', *Computational Intelligence and Neuroscience* **2017**, 1–11.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A. & Carin, L. (2016), Variational autoencoder for deep learning of images, labels and captions, *in* 'Advances in Neural Information Processing Systems', pp. 2352–2360.

Puelz, R. & Kemmsies, W. (1993), 'Implications for unisex statutes and risk-pooling: The costs of gender and underwriting attributes in the automobile insurance market', *Journal of Regulatory Economics* **5**(3), 289–301.

Puelz, R. & Snow, A. (1994), 'Evidence on adverse selection: Equilibrium signaling and cross-subsidization in the insurance market', *Journal of Political Economy* **102**(2), 236–257.

Quinlan, J. R. (2014), *C4. 5: Programs for machine learning*, Elsevier.

Ramanathan, R., Engle, R., Granger, C. W., Vahid-Araghi, F. & Brace, C. (1997), 'Short-run forecasts of electricity loads and peaks', *International Journal of Forecasting* **13**(2), 161–174.

Ravallion, M., Galasso, E., Lazo, T. & Philipp, E. (2005), 'What can ex-participants reveal about a program's impact?', *Journal of Human Resources* **40**(1), 208–230.

Reguant, M. (2014), 'Complementary bidding mechanisms and startup costs in electricity markets', *The Review of Economic Studies* **81**(4), 1708–1742.

Reus, L., Munoz, F. D. & Moreno, R. (2018), 'Retail consumers and risk in centralized energy auctions for indexed long-term contracts in Chile', *Energy policy* **114**, 566–577.

Richaudeau, D. (1999), 'Automobile insurance contracts and risk of accident: An empirical test using French individual data', *The Geneva Papers on Risk and Insurance Theory* **24**(1), 97–114.

Rotemberg, J. & Saloner, G. (1986), 'A supergame-theoretic model of price wars during booms', *New Keynesian Economics* **2**, 387–415.

Rothschild, M. & Stiglitz, J. E. (1976), 'Increasing risk: A definition', *Journal of Economic Theory* **2**, 225–243.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Rousseeuw, P. J. & Driessen, K. V. (1999), 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**(3), 212–223.

Ruder, S. (2016), 'An overview of gradient descent optimization algorithms', *arXiv preprint arXiv:1609.04747* .

Saito, K. (2006), 'Testing for asymmetric information in the automobile insurance market under rate regulation', *Journal of Risk and Insurance* **73**(2), 335–356.

Saitta, S., Raphael, B. & Smith, I. F. (2007), A bounded index for cluster validity, *in* 'International Workshop on Machine Learning and Data Mining in Pattern Recognition', Springer, pp. 174–187.

Samson, D. (1986), 'Designing an automobile insurance classification system', *European Journal of Operational Research* **27**(2), 235–241.

Samson, D. & Thomas, H. (1987), 'Linear models as aids in insurance decision making: the estimation of automobile insurance claims', *Journal of Business Research* **15**(3), 247–256.

Santosa, F. & Symes, W. W. (1986), 'Linear inversion of band-limited reflection seismograms', *SIAM Journal on Scientific and Statistical Computing* **7**(4), 1307–1330.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. (2001), 'Estimating the support of a high-dimensional distribution', *Neural Computation* **13**(7).

Schreyer, M., Sattarov, T., Borth, D., Dengel, A. & Reimer, B. (2017), 'Detection of anomalies in large scale accounting data using deep autoencoder networks', *arXiv preprint arXiv:1709.05254* .

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Sculley, D. (2010), Web-scale k-means clustering, *in* 'Proceedings of the 19th International Conference on World Wide Web', ACM, pp. 1177–1178.

Seaborn, C., Attanucci, J. & Wilson, N. H. (2009), 'Using smart card fare payment data to analyze multi-modal public transport journeys in London', *Journal of the Transportation Research Board* **2121**, 55–62.

Sekuła, P., Marković, N., Vander Laan, Z. & Sadabadi, K. F. (2018), 'Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study', *Transportation Research Part C: Emerging Technologies* **97**, 147–158.

Shmueli, G. et al. (2010), 'To explain or to predict?', *Statistical Science* **25**(3), 289–310.

Skrzypacz, A. & Hopenhayn, H. (2004), 'Tacit collusion in repeated auctions', *Journal of Economic Theory* **114**(1), 153–169.

Smith, K. A., Willis, R. J. & Brooks, M. (2000), 'An analysis of customer retention and insurance claim patterns using data mining: A case study', *Journal of the Operational Research Society* **51**(5), 532–541.

Soares, L. J. & Medeiros, M. C. (2005), Modelling and forecasting short-term electricity load: a two step methodology, Technical report, Department of Economics PUC-Rio (Brazil).

Soares, L. J. & Souza, L. R. (2006), 'Forecasting electricity demand using generalized long memory', *International Journal of Forecasting* **22**(1), 17–28.

Spence, M. (1978), Job market signaling, *in* 'Uncertainty in Economics', Elsevier, pp. 281–306.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014), 'Dropout: a simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research* **15**(1), 1929–1958.

Stefano, B. & Gisella, F. (2001), Insurance fraud evaluation: a fuzzy expert system, *in* 'Fuzzy Systems, 2001. The 10th IEEE International Conference on', Vol. 3, IEEE, pp. 1491–1494.

Stiglitz, J. E. (1977), 'Monopoly, non-linear pricing and imperfect information: the insurance market', *The Review of Economic Studies* **44**(3), 407–430.

Strbac, G. & Wolak, F. A. (2017), Electricity market design and renewables integration in developing countries, Technical report.

Suwardo, W., Napiah, M. & Kamaruddin, I. (2010), 'ARIMA models for bus travel time prediction', *Journal of the Institute of Engineers Malaysia* pp. 49–58.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2818–2826.

Szkuta, B., Sanabria, L. A. & Dillon, T. S. (1999), 'Electricity price short-term forecasting using artificial neural networks', *IEEE Transactions on Power Systems* **14**(3), 851–857.

Tan, C. C. & Eswaran, C. (2010), 'Reconstruction and recognition of face and digit images using autoencoders', *Neural Computing and Applications* **19**(7), 1069–1079.

Tan, Z., Zhang, J., Wang, J. & Xu, J. (2010), 'Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models', *Applied Energy* **87**(11), 3606–3610.

Tao, S., Corcoran, J., Hickman, M. & Stimson, R. (2016), 'The influence of weather on local geographical patterns of bus usage', *Journal of Transport Geography* **54**, 66–80.

Tao, S., Rohde, D. & Corcoran, J. (2014), 'Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap', *Journal of Transport Geography* **41**, 21–36.

Tefft, B. C. (2008), 'Risks older drivers pose to themselves and to other road users', *Journal of Safety Research* **39**(6), 577–582.

Theis, L., Shi, W., Cunningham, A. & Huszár, F. (2017), 'Lossy image compression with compressive autoencoders', *arXiv preprint arXiv:1703.00395* .

Tikhonov, A. N. & Arsenin, V. I. (1977), *Solutions of ill-posed problems*, Vol. 14, Vh Winston.

Tirachini, A., Hensher, D. A. & Rose, J. M. (2013), 'Crowding in public transport systems: effects on users, operation and implications for the estimation of demand', *Transportation Research Part A: Policy and Practice* **53**, 36–52.

Trépanier, M., Morency, C. & Agard, B. (2009), 'Calculation of transit performance measures using smartcard data', *Journal of Public Transportation* **12**(1), 5.

Trivedi, S., Pardos, Z. A. & Heffernan, N. T. (2015), 'The utility of clustering in prediction tasks', *arXiv preprint arXiv:1509.06163* .

Tryfos, P. (1980), 'On classification in automobile insurance', *The Journal of Risk and Insurance* **47**(2), 331–337.

Tsai, C.-H., Mulley, C. & Clifton, G. (2014), 'A review of pseudo panel data approach in estimating short-run and long-run public transport demand elasticities', *Transport Reviews* **34**(1), 102–121.

Tsai, C.-H. P. & Mulley, C. (2014), 'Identifying short-run and long-run public transport demand elasticities in Sydney a pseudo panel approach', *Journal of Transport Economics and Policy (JTEP)* **48**(2), 241–259.

Tumay, M. (2009), 'Asymmetric information and adverse selection in insurance markets: the problem of moral hazard', *Yönetim ve Ekonomi* **16**(1), 107–114.

Utsunomiya, M., Attanucci, J. & Wilson, N. (2006), 'Potential uses of transit smart card registration and transaction data to improve transit planning', *Transportation Research Record: Journal of the Transportation Research Board* (1971), 119–126.

Van Der Maaten, L., Postma, E. & Van den Herik, J. (2009), 'Dimensionality reduction: a comparative', *Journal of Machine Learning Research* **10**(66-71), 13.

Varian, H. R. (2014), 'Big data: New tricks for econometrics', *Journal of Economic Perspectives* **28**(2), 3–28.

Vellido, A., Martín-Guerrero, J. D. & Lisboa, P. J. (2012), Making machine learning models interpretable, *in* 'ESANN', Vol. 12, Citeseer, pp. 163–172.

Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D. & Dedene, G. (2007), 'Strategies for detecting fraudulent claims in the automobile insurance industry', *European Journal of Operational Research* **176**(1), 565–583.

Villaplana Conde, P., Peña Sánchez de Rivera, J. I., Escribano Sáez, Á. et al. (2002), Modeling electricity prices: international evidence, Technical report, Universidad Carlos III de Madrid. Departamento de Economía.

von der Fehr, N.-H. M. & Harbord, D. (1992), *Long-term contracts and imperfectly competitive spot markets: a study of the UK electricity industry*, Univ., Department of Economics.

von der Fehr, N.-H. M. & Harbord, D. (1993), 'Spot market competition in the UK electricity industry', *The Economic Journal* **103**(418), 531–546.

Walker, J., Doersch, C., Gupta, A. & Hebert, M. (2016), An uncertain future: Forecasting from static images using variational autoencoders, *in* 'European Conference on Computer Vision', Springer, pp. 835–851.

Wang, D., Luo, H., Grunder, O., Lin, Y. & Guo, H. (2017), 'Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm', *Applied Energy* **190**, 390–407.

Wang, F. & Ross, C. L. (2018), 'Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model', *Transportation Research Record* **2672**(47), 35–45.

Wang, J., Chen, R. & He, Z. (2019), 'Traffic speed prediction for urban transportation network: A path based deep learning approach', *Transportation Research Part C: Emerging Technologies* **100**, 372–385.

Wang, Y., Zhang, D., Liu, Y., Dai, B. & Lee, L. H. (2019), 'Enhancing transportation systems via deep learning: A survey', *Transportation Research Part C: Emerging Technologies* **99**, 144–163.

Wang, Z. J., Li, X. H. & Chen, F. (2015), 'Impact evaluation of a mass transit fare change on demand and revenue utilizing smart card data', *Transportation Research Part A: Policy and Practice* **77**, 213–224.

Weiss, G. M. (2004), 'Mining with rarity: a unifying framework', *ACM SIGKDD Explorations Newsletter* **6**(1), 7–19.

Weron, R. (2014), 'Electricity price forecasting: A review of the state-of-the-art with a look into the future', *International journal of forecasting* **30**(4), 1030–1081.

Weron, R. & Misiorek, A. (2008), 'Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models', *International Journal of Forecasting* **24**(4), 744–763.

Williams, B. M., Durvasula, P. K. & Brown, D. E. (1998), 'Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models', *Transportation Research Record* **1644**(1), 132–141.

Williams, G. J. (1999), Evolutionary hot spots data mining, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining', Springer, pp. 184–193.

Williams, G. J. & Huang, Z. (1997), Mining the knowledge mine, *in* 'Australian Joint Conference on Artificial Intelligence', Springer, pp. 340–348.

Wilson, C. (1977), 'A model of insurance markets with incomplete information', *Journal of Economic theory* **16**(2), 167–207.

Wilson, J. H. (2009), 'An analytical approach to detecting insurance fraud using logistic regression', *Journal of Finance and Accountancy* **1**, 1.

Wolak, F. A. (2000), 'An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive electricity market', *International Economic Journal* **14**(2), 1–39.

Wolak, F. A. (2017), Measuring the impact of purely financial participants on wholesale and retail market performance: The case of Singapore, Technical report, Department of Economics, Stanford University.

Woo, C. K., Horowitz, I. & Hoang, K. (2001), 'Cross hedging and forward-contract pricing of electricity', *Energy Economics* **23**(1), 1–15.

Woo, C. K., Horowitz, I., Horii, B. & Karimov, R. I. (2004), 'The efficient frontier for spot and forward purchases: an application to electricity', *Journal of the Operational Research Society* **55**(11), 1130–1136.

Worthington, A., Kay-Spratley, A. & Higgs, H. (2005), 'Transmission of prices and price volatility in australian electricity spot markets: a multivariate GARCH analysis', *Energy Economics* **27**(2), 337–350.

Wright, P. G. (1928), *Tariff on animal and vegetable oils*, Macmillan Company, New York.

Wu, Y., Tan, H., Qin, L., Ran, B. & Jiang, Z. (2018), 'A hybrid deep learning based traffic flow prediction method and its understanding', *Transportation Research Part C: Emerging Technologies* **90**, 166–180.

Xiao, L., Shao, W., Yu, M., Ma, J. & Jin, C. (2017), 'Research and application of a hybrid wavelet neural network model with the improved cuckoo search algorithm for electrical power system forecasting', *Applied energy* **198**, 203–222.

Xie, C., Lu, J. & Parkany, E. (2003), 'Work travel mode choice modeling with data mining: decision trees and neural networks', *Transportation Research Record* **1854**(1), 50–61.

Xu, C., Ji, J. & Liu, P. (2018), 'The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets', *Transportation Research Part C: Emerging Technologies* **95**, 47–60.

Xue, R., Sun, D. J. & Chen, S. (2015), 'Short-term bus passenger demand prediction based on time series model and interactive multiple model approach', *Discrete Dynamics in Nature and Society* .

Yang, Y., Qian, W. & Zou, H. (2015), 'A boosted tweedie compound poisson model for insurance premium', *arXiv preprint arXiv:1508.06378* .

Yang, Z., Ce, L. & Lian, L. (2017), 'Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods', *Applied Energy* **190**, 291–305.

Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K. & Bethard, S. (2013), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, *in* 'Conference on Empirical Methods in Natural Language Processing'.

Yeo, A. C., Smith, K. A., Willis, R. J. & Brooks, M. (2001), 'Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry', *Intelligent Systems in Accounting, Finance & Management* **10**(1), 39–50.

Yi, D., Su, J., Liu, C., Quddus, M. & Chen, W.-H. (2019), 'A machine learning based personalized system for driving state recognition', *Transportation Research Part C: Emerging Technologies* **105**, 241–261.

Yu, B., Lam, W. H. & Tam, M. L. (2011), 'Bus arrival time prediction at bus stop with multiple routes', *Transportation Research Part C: Emerging Technologies* **19**(6), 1157–1170.

Yu, L., Wang, S. & Lai, K. K. (2008), 'Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm', *Energy Economics* **30**(5), 2623–2635.

Yu, L., Zhao, Y. & Tang, L. (2014), 'A compressed sensing based AI learning paradigm for crude oil price forecasting', *Energy Economics* **46**, 236–245.

Yuan, F. & Cheu, R. L. (2003), 'Incident detection using support vector machines', *Transportation Research Part C: Emerging Technologies* **11**(3-4), 309–328.

Zhai, S., Cheng, Y., Lu, W. & Zhang, Z. (2016), 'Deep structured energy based models for anomaly detection', *arXiv preprint arXiv:1605.07717* .

Zhang, Y. & Xie, Y. (2008), 'Travel mode choice modeling with support vector machines', *Transportation Research Record* **2076**(1), 141–150.

Zhang, Y., Yao, E., Zhang, J. & Zheng, K. (2018), 'Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data', *Transportation Research Part C: Emerging Technologies* **92**, 76–89.

Zhao, X., Yan, X., Yu, A. & Van Hentenryck, P. (2018), 'Modeling stated preference for mobility-on-demand transit: A comparison of machine learning and logit models', *arXiv preprint arXiv:1811.01315* .

Zhao, Z., Koutsopoulos, H. N. & Zhao, J. (2018), 'Individual mobility prediction using transit smart card data', *Transportation Research Part C: Emerging Technologies* **89**, 19–34.

Zhou, C. & Paffenroth, R. C. (2017), Anomaly detection with robust deep autoencoders, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, New York, NY, USA, pp. 665–674.

Zhou, D., Bousquet, O., Navin Lal, T., Weston, J. & Scholkopf, B. (2004), 'Learning with local and global consistency', *In Advances in neural information processing systems* pp. 321–328.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W. & Cao, R. (2017a), 'Impacts of weather on public transport ridership: Results from mining data from different sources', *Transportation Research Part C: Emerging Technologies* **75**, 17–29.

Zhou, M., Yan, Z., Ni, Y., Li, G. & Nie, Y. (2006), 'Electricity price forecasting with confidence-interval estimation through an extended ARIMA approach', *IEE Proceedings-Generation, Transmission and Distribution* **153**(2), 187–195.

Zhou, Y., Yao, L., Chen, Y., Gong, Y. & Lai, J. (2017b), 'Bus arrival time calculation model based on smart card data', *Transportation Research Part C: Emerging Technologies* **74**, 81–96.

Zhu, X. & Ghahramani, Z. (2002), Learning from labeled and unlabeled data with label propagation, Technical report, CMU-CALD-02-107, Carnegie Mellon University.