BACHELOR'S DEGREE THESIS

Degree in Statistics

Title: Approaching the cold start problem in customer relationship management through lifetime value models.

Author: Daniel Guivernau Rosés

Advisor: Lourdes Rodero De Lamo

Department: Statistics and Operations Research

Academic year: 2021-2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Abstract

Companies are often interested in assessing the future profitability of their customers. This is especially challenging for clients that the company recently acquired, given the small amount of information available.

In this thesis, we investigate a different way of approaching this problem, using the Pareto/NBD and Gamma/Gamma models. These allow us to establish statistical descriptions of the purchase rates, the length of the relation with a client and the average value of their purchases.

In the end with these models we will create a binary classifier that, given the purchase history of a group of clients acquired at a certain time will tell us which are expected to be profitable in the future, which will present extremely high specificity and acceptable sensibility.

Keywords: Customer relationship management, Cold start problem, Pareto/NBD, Empirical Bayesian models, Classification models.

Resum

En el món empresarial és d'interès avaluar la rendibilitat futura dels clients. Això és particularment difícil per a clients que l'empresa acaba de captar, ja que se'n té menys informació.

En aquesta tesi investiguem una manera diferent d'afrontar aquest problema, utilitzant els models Pareto/NBD i Gamma/Gamma. Aquests ens permeten establir descripcions estadístiques de la freqüència de compra, la durada de la relació amb un client i el valor mitjà de les seves compres.

Emprant aquests models al final crearem un classificador binari que, donada l'historial de compres d'un grup de clients en un determinat període de temps, ens dirà quins seran rendibles en el futur, i que presentarà molt bona especificitat i sensibilitat acceptable.

Paraules clau: Gestió de relació amb clients, El problema d'arrencada en fred, Pareto/NBD, Models Bayesians empírics, Models de classificació.

Table of contents.

1. Introduction

Broadly speaking, a Catalan company that sells hardware equipment is looking for a statistical way of determining whether a client will be profitable in the long run upon their first purchases. In the field of marketing this is known as the Cold start problem, and there are a multitude of models and research papers dedicated to tackling it.

There's been substantial work done by other university students on this particular dataset, and both logistic regressions and more sophisticated statistical learning algorithms have been employed. This thesis aims to explore alternative ways of tackling this problem to hopefully rectify some of the issues that were encountered in previous efforts.

To do this we will employ the Pareto/NBD and the Gamma/Gamma spending models to approximate the probabilistic processes behind the frequency of purchases, the length of the relationship between the company and its clients, and the value of the purchases, so that by the end we can construct a classifier that tells us whether a client will be good in the future.

This thesis will be organized as follows. In section 2 we will give a broad introduction into the models we will be using, what assumptions they make, what results we can obtain from them, and the software needed to fit these models. In section 3 we adjust the models, going from a short descriptive analysis of our data to the validation of the models' predictions. Finally, in section 4 we take the predictions these models allow us to make, translate them into the final client classifier and analyze its precision.

2. Methodology

2.1. Situation

The object of our research is to construct a statistical model that after the first purchases of a new client will classify them into either good or bad, where good stands for profitable in the future. Having said that, the company in question gave a very specific definition of profitable, and a client may only be considered good if:

- The client is alive: they made at least one purchase in the last year.
- On average, the client has spent 1000€ per year.

As we can see, this definition naturally translates into a binary response variable. Therefore, logistic regressions were the approach taken by almost all other previous students: on the last day of the studied period every client was either good or bad, and variables such as the value and the variability of the first purchase and what delegation the client was registered in were supposed to predict this binary variable.

Independently of model specification, there are certain issues that jump to mind when considering this approach, for example:

- If a client has only purchased once, in the last studied year, and spent over a thousand euros, this client is automatically good, its response variable is perfectly determined by its first purchase value.
- Considering that data from 2010 to 2020 is available (even if only the period 2012-2019 was studied, for reasons we will explain bellow) a client may have been good for most of the period and then stopped purchasing on the last year. This client is then classified as bad, despite having been a very good client for over 8 years.
- This model also assumes a client only dies once. Since the life status of the client is based on recency, a client could have purchased once in 2012, and then again in 2019, and they would be considered alive, despite having gone on a seven year purchase hiatus.

On top of that, there were other problems with the concrete specification of the models:

- Purchase value was included both as a predictive variable and as part of what determined the response variable, a practice that is generally thought of as improper, as it can inflate model bias.
- Many of the predictive variables were categorical and with a large amount of levels, and in certain mixed models these were shown to be unfairly significant, even in simulated situations where they shouldn't have been. For more information, see Carbonell, 2020.

2.2. The cold start problem

The problem described above is known in marketing, and more precisely in the field of customer relationship management, as the cold start problem, and it exists and has been studied in every possible field of business and through multiple disciplines.

When it comes to statistical tools to approach this problem, models based on a customer's lifetime value are perhaps some of the most popular. This metric is defined as the profit a customer provided the company throughout their whole commercial relationship. It therefore quantifies how "useful" a customer has been for the company. If we can calculate this metric, applying a generalized linear model or machine learning tools to see what information in a customer's first purchases is related to the lifetime value gives us a solid way of overcoming the cold start problem.

Having said that, most of the ways of calculating a customer's lifetime value are built for contractual relationships, that is, fields of business where there is a precise time in which the relationship between company and customer comes to an end. Think, for example, legal services or utility providers.

This isn't our situation. Since our company is a retailer, a customer does not notify when they intend to stop buying. And therefore, the lifetime duration of each customer is completely unknown to us. Having said that, there is a family of empirical Bayesian models created with our situation in mind. These are the Pareto/NBD and the Gamma/Gamma models, which we will describe in detail in the following sections.

2.3. The Pareto/NBD model

The first formulation for the Pareto/NBD model was presented in “Counting Your Customers: Who Are They and What Will They Do Next?” (Schmittlein, Morrison, & Colombo, 1987).

This model is concerned with evaluating existing clients of a business with non-contractual relationships. More precisely, given a group of clients acquired at a certain time, it aims to estimate how likely each customer is to still be active (we will call active clients “alive” and inactive ones “dead”), and what amount of transactions can be expected of them in the future.

In order to do this, for every client we must know:

- How long they were observed for, T: the date we stop collecting data minus the date of their first purchase, measured in whatever granularity we're interested in or that our data allows.
- How many purchases they made during that time, X.
- The time of their last purchase t, which must validate $0 \leq t \leq T$.

Because this model is intended for non-contractual relationships, it will estimate both the distribution the purchase rate while a client is alive, and the probability that the client has died. We will make the following assumptions about these processes:

- Alive clients make purchases following a Poisson process with rate λ.
- The time each client is alive follows an exponential distribution with death or attrition rate μ.
- The values of λ for different customers follow a Gamma distribution with α as the rate parameter and r as the shape parameter.
- The values of μ for different customers follow a Gamma distribution with β as the rate parameter and s as the shape parameter.
- The variables μ and λ are independent.

Of course, it is possible to have data that that doesn't meet these assumptions. The original authors argue that for frequently purchased goods the Poisson distribution for the purchases is reasonable while this isn't the case for catalog purchases, as the creation and distribution of new catalogs will probably interfere with the constant daily probability of purchasing that our suppositions imply. We must keep this in mind, as it will become important in our own data later on.

Given these assumptions and their mathematical formulation, we are now interested in predicting the probability of being alive and the expected amount of future purchases for every client. Although the key concepts and formulations will be presented here, we suggest that readers interested in the precise mathematical derivation of these results direct themselves to the original paper.

Firstly, if we denote $\tau > 0$ as the unobservable time at which the client died, we can find: $\Pr(X = x|\, \lambda, \tau > T) = \frac{e^{-\lambda T}(\lambda T)^x}{x!}$, where x is a natural number and $f(\tau|\mu) = \mu e^{-\mu\tau}$, which, after some calculations, allow us to derive the probability that the client is alive at the end of the observation period:

$$\Pr(\tau > T \mid \lambda, \mu, X = x, t, T) = \left(1 + \frac{\mu}{\lambda+\mu}(e^{(\lambda+\mu)(T-t)} - 1)\right)^{-1}$$

Of course, although theoretically interesting, this formula has no practical use, as the quantities μ and λ are unknown. And thus, our objective will be to estimate their distributions over the whole customer base so that we can later on calculate that probability.

We know that those distributions will have the form $f(\lambda|r,\alpha) = \frac{\alpha^r}{\Gamma(r)}\lambda^{r-1}e^{-\alpha\lambda}$, where $\lambda, r, \alpha > 0$ and therefore $E(\lambda|r,\alpha) = \frac{r}{\alpha}$ and $Var(\lambda|r,\alpha) = \frac{r}{\alpha^2}$ and analogously, $f(\mu|s,\beta) = \frac{\beta^s}{\Gamma(s)}\mu^{s-1}e^{-\beta\mu}$, where $\mu, s, \beta > 0$ and therefore $E(\mu|s,\beta) = \frac{s}{\beta}$ and $Var(\mu|s,\beta) = \frac{s}{\beta^2}$.

All that we've seen before and the assumptions we made lead us to the complete specification of the model. Firstly, while the customer is alive their purchases follow the NBD model such that:

$$Pr(X = x \mid r,\alpha,\tau > T) = \binom{x+r-1}{x}\frac{\alpha^r}{(\alpha+T)^r}\frac{T^x}{(\alpha+T)^x}$$, where x is a natural number.

And the time of death follows a type 2 Pareto distribution such that:

$$f(\tau|s,\beta) = \frac{s}{\beta}\left(\frac{\beta}{\beta+\tau}\right)^{s+1}$$, where $\tau > 0$.

These are the distributions that inform the model's name, and we can see that they don't depend on neither λ nor μ. It is also the presence of the Negative binomial distribution that will make the estimation of this model computationally complicated, as we'll see once we fit it to our data.

From this point onwards we'll present the mathematical formulation of the main results we'll get out of this model, the metrics we are interested in. These aren't included in order to understand them deeply, just mentioned as we will use them in our application.

The first result we're interested in is the probability of being alive at the end of the estimation period, which is calculated as follows:

If $\alpha > \beta$,

$$Pr(\tau > T \mid r,s,\alpha > \beta, X = x,t,T) = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\alpha+t}\right)^{r+x}\left(\frac{\beta+T}{\alpha+t}\right)^{s}F(a,b_1;c;z_1(t)) - \left(\frac{\beta+T}{\alpha+T}\right)^{s}F(a,b_1;c;z_1(t))\right]\right\}^{-1}$$

If $\alpha < \beta$,

$$Pr(\tau > T \mid r,s,\alpha < \beta, X = x,t,T) = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\beta+t}\right)^{r+x}\left(\frac{\beta+T}{\beta+t}\right)^{s}F(a,b_2;c;z_2(t)) - \left(\frac{\alpha+T}{\beta+T}\right)^{r+x}F(a,b_2;c;z_2(t))\right]\right\}^{-1}$$

If $\alpha = \beta$,

$$\Pr(\tau > T \mid r, s, \alpha = \beta, X = x, t, T) = \left\{1 + \frac{s}{r+x+s}\left[\left(\frac{\alpha+T}{\alpha+t}\right)^{r+x+s} - 1\right]\right\}^{-1}$$

Where $a = r + x + s$, $b_1 = s + 1$, $b_2 = r + x$, $c = r + x + s + 1$, $z_1 = \frac{\alpha-\beta}{\alpha+y}$, $z_2 = \frac{\beta-\alpha}{\beta+y}$ and $F(a, b; c; z)$ is the Gauss hypergeometric function.

Another result we're interested in is the expected amount of purchases for a randomly chosen customer during the estimation period, which can be calculated as $E(X|r, \alpha, s, \beta, T) = \frac{r\beta}{\alpha(s-1)}\left[1 - \left(\frac{\beta}{\beta+T}\right)^{s-1}\right]$.

The final calculation we're interested in, and the one that will play the most important role in out particular use of this model, is the expected amount of purchases in a future time period. Knowing that a customer made X purchases in the estimation period $(0, T]$, how many are they expected to make in $(T, T + T^*]$?

Given the nice properties of the distributions we're working with, this result isn't hard to find. First we see that a client such that $\tau < T$ will necessarily make zero purchases after T. Otherwise, the customer will still follow a Pareto/NBD process with the updated parameters $r^* = r + x$, $\alpha^* = \alpha + T$, $s^* = s$, $\beta^* = \beta + T$, due to the lack of memory of the underlying processes in this model.

Using this we get to the expected value $E(X^*|r, \alpha, s, \beta, X = x, t, T, T^*) = E(X^*|r + x, \alpha + T, s, \beta + T, T^*)\Pr(\tau > T|r, \alpha, s, \beta, X = x, t, T)$

What we now described is the simple Pareto/NBD model. With it we can properly describe, predict and evaluate the amount of purchases a certain group of clients will make in the future. Having said that, because of the object of our research we are also interested in an extended version of this model which allows us to include a set of static covariates. We intend to include information obtained during the first purchase of a client as these covariates in order to make this model more precise and to evaluate the effects of these variables.

This extended model is described in "Incorporating Time-Invariant Covariates into the Pareto/NBD and BG/NBD Models" by Peter S. Fader and Bruce G.S. Hardie.

The idea is simple, we replace α and β with the expressions $\alpha = \alpha_0 e^{-\gamma'_1 z_1}$ and $\beta = \beta_0 e^{-\gamma'_2 z_2}$, where z_1 and z_2 are the vector of static covariates that we think explain some of the variance in the purchasing process and the dropout process respectively (in our case, $z_1 = z_2$) and γ_1 and γ_2 are their respective coefficient vectors.

Afterwards, although the computational adjustment of the model is slightly different, we'll obtain the same results, and additionally, we will be able to see if every coefficient significantly alters the purchase and dropout processes, and how they do this.

The interpretation of the coefficients will be the same we would make in a generalized linear model. For example, suppose we include the binary variable sex into the model, taking values male and female, with male being the reference category. If the dummy variable's coefficient in the transaction process is positive, then females have an expected higher purchase rate than males, while if it's negative, their expected purchase rate is lower than that of males. If the variable's coefficient in the lifetime process is positive, females would be expected to be active customers for longer than males, while if it's negative, they would be expected to be active customers for a lower amount of time than males.

With this, we've gotten to the point where we can describe the purchase and dropout processes of clients accurately and taking into account the information we obtained during their first purchases. Although we are getting closer to our objective, we are still missing a crucial piece, we must be able to also estimate how valuable the purchases of every client will be. We will do this using the following model, the Gamma/Gamma spending model.

2.4. The Gamma/Gamma spending model

In this section we will describe the model we will use to estimate how valuable the average purchase of every client will be, which, combined with the previous model, will allow us to accurately estimate the CLV of every client.

This model is presented in detail in Fader & Hardie 2013, here we will summarize its assumptions, parameters and main results.

The assumptions this model is based on are:

- The monetary value of a customer's transaction varies around their average transaction value, the quantity we want to estimate.
- The average transaction value of a customer does not vary over time, but it does vary across the customer base.
- The average transaction value of a customer is independent of their transaction and lifetime process.

If a customer has x transactions and $z_1, z_2, \dots, z_x$ are their values we can estimate their mean $\bar{z}$, which is an estimate of the population mean we want to estimate, $E(Z|\, \bar{z}, x)$.

We suppose that:

- Z_i follows a Gamma distribution with shape parameter p and rate parameter v, such that $E(Z_i|\, p, v\,) = \frac{p}{v}$ and $\bar{z}$ follows a gamma distribution with shape and rate parameters px and vx, respectively.
- v follows a Gamma distribution with shape parameter q and rate parameter γ.

After finding the values of the parameters by maximizing the log-likelihood function we can calculate:

$$E(Z \mid p, q, \gamma; \bar{z}, x) = \frac{p(\gamma + x\bar{z})}{px + q - 1}$$

Which is the expected average purchase value of every customer, the quantity we need to calculate their expected CLV.

2.5. The CLVtools library

Both of the presented models will be applied in our thesis using the R programming language and the package CLVtools.

The general workflow of this package will be as follows. First we will use a database of the purchase times and values and a separate one of covariates to create a particular data object, which will be used later for model fitting. When creating this object we will specify the granularity of our data and what period of time will be used for model adjustment, so that the rest can be left for model validation.

After fitting the Pareto/NBD model we will obtain an S4 object we can use to plot fitted quantities and to make predictions. We will make predictions about our clients over the validation period so that we can see how well our model predicts the amount of purchases, average purchase value and probability of being alive for every client.

3. Model creation and validation

3.1. Available data

The data that we will use in this thesis comes in the following forms: a database of purchase tickets that ranges from 2010 to 2020 and a database containing information about the clients. Note however that these aren't the original databases obtained from the business, but rather the ones used by the previous students mentioned before. Therefore, these have already been treated and have had several issues corrected.

The first database includes the following variables:

- CustomerId: Identification code of the client.
- n_container: Purchase identifier. An individual purchase may contain multiple tickets and therefore this identifier is necessary.
- Date_delivery: Date on which the transaction was registered.
- Up, family and sku: Levels of product identification. Inside every up group there are multiple families, and inside every family there's different skus, which are product identification numbers.
- Qty: Quantity of skus purchased in that line
- Price: Unitary price of that line's sku, measured in Euros.
- Disc: Discount applied to that line, measured in percentages.
- Bill: Total value of that line, calculated as qty*price - discount.
- Descshipping: Whether the purchase was shipped or picked up by the client.
- Store: Whether the purchase was made in person at the company's physical store or through other means.
- Treatment: Categorical variable that takes either “Silver” or “Gold”. It identifies how valuable the company thought the client was at the moment of purchase.
- Digital client: Whether the client registered through the company's website

The client information database includes the following variables:

- CustomerId: Identification code of the client
- Delegation: Through which delegation the client first contacted the company.
- Country: What country the client is in.

- Market: what market the business classified the client into, can take the following values: MRO (maintenance and repair), OEM (original equipment manufacturing), SI (industrial systems) and OTR (other).
- Province: What province of Spain the client is in, takes the value “international” if the client is from abroad.

3.2. Data cleaning and creation of covariates

With the data described above, we set out to create a set of variables that codify all the information the company would have upon a customer's first purchase that we think could contain valuable information.

The first thing we did was to eliminate the tickets after 2019-12-31, as the pandemic that later started out in early 2020 most likely had an impact on the statistical processes we intend to estimate that we can't possibly account for in this thesis.

We also decided to eliminate international clients from the data, as one of the variables we wanted to calculate, the distance between the province of the client and the delegation they were assigned to couldn’t be calculated for these clients and an artificial value had to be inputted. We could afford this loss of data, as only 115 out of 21603 clients were international.

Afterwards we used all the databases available to us to create a covariate dataset, which, with one line per client, would include the following information about said client and their first purchase:

Firstly, information about the client:

- CustomerId, delegation, province and market as described above.
- Same zone: binary variable that takes the value 1 if the delegation and province variables take the same value and 0 otherwise.
- Distance: distance between the delegation and province, taken from a fixed distance matrix.

Secondly, information about the first purchase:

- First purchase and First purchase month: Date of the first purchase the customer ever made and month of that date.
- Transport: Whether the first purchase was picked up by the client or delivered through other means.

- Treatment, store and digital client: values of these variables described above in the customer's first purchase.
- First purchase lines: Number of lines the first purchase was made up of, that is, how many different items or skus it included.
- First purchase value: Total monetary value of all the lines the first purchase was made up of.

Since the methods we will use to adjust the Pareto model require that no covariate has missing values, we also removed all the clients that had any. This wasn't a big problem, as the amount of clients went from 21488 to 21316.

So we can better illustrate our situation, let's do a brief descriptive analysis of this new covariate database.

3.3. Descriptive analysis

We begin by looking at the first purchase dates of our database. As we can see in figure 3.1, there appear to be two different periods.

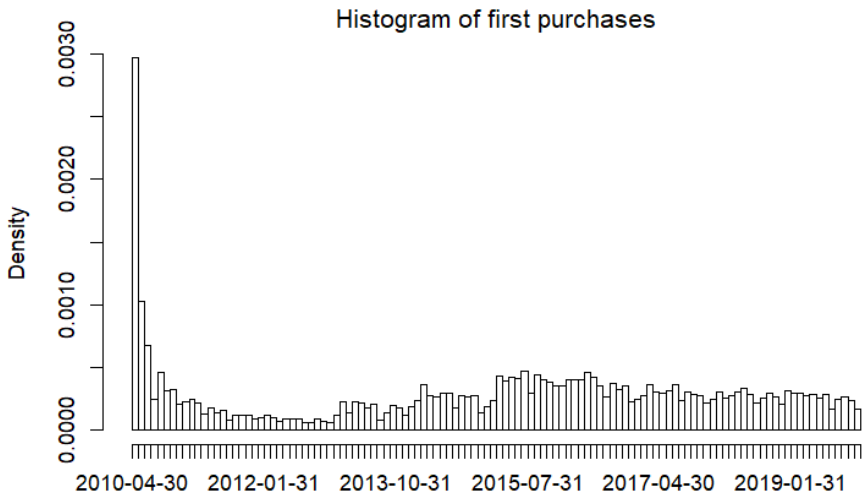


Figure 3.1, histogram of first purchases with monthly breaks

First, there's a very large number of new clients on the first months of 2010, which decreases steadily until late 2012. We must keep in mind that although this database has been kept since 2010, the company had been operating for much longer, so we have reason to believe that those aren't new clients but the first purchases of existing clients after data collection had started.

Later, from 2013 to 2016 the amount of first purchases steadily increases and then decreases slightly until the end of 2019. This can be seen more clearly in figure 3.2.

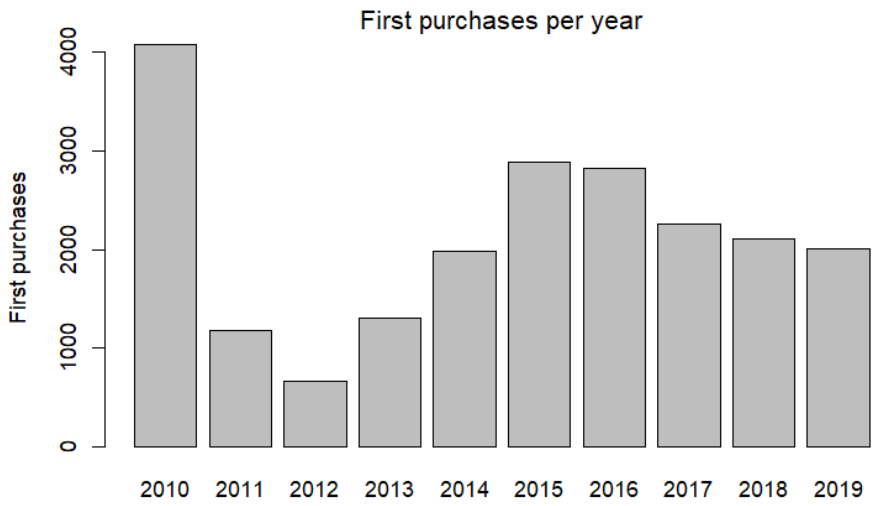


Figure 3.2, bar plot of first purchases per year

When it comes to the months when the first purchases were made, as we can see in figure 3.3, they aren't evenly distributed. May has the highest value, slightly above 3500 while August and December are barely above 1000. This is probably due to holiday seasons and different levels of economic activity throughout the year.

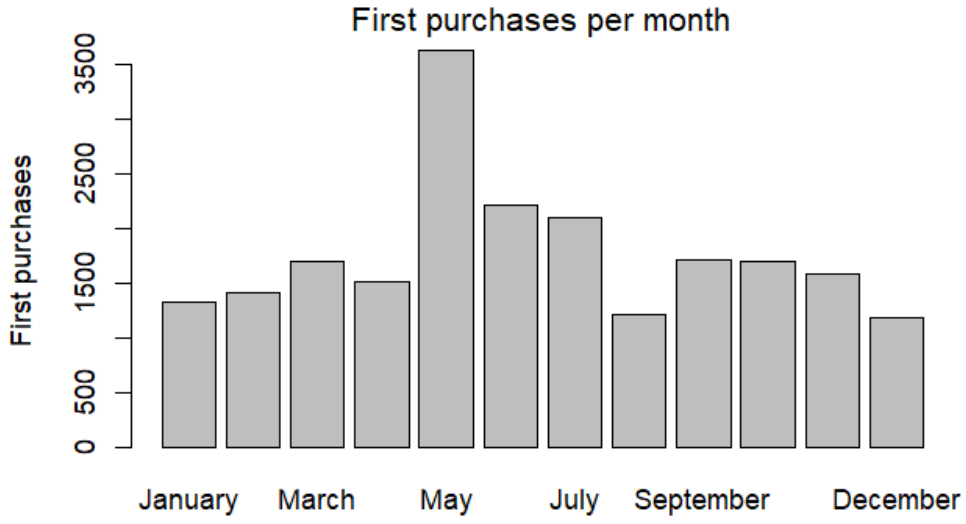


Figure 3.3, bar plot of first purchases per month

Moving on to the values and number of lines of the first purchases, we can see in table 3.1 that both are extremely asymmetrical and right-skewed. We can see they have a median of 2 lines and 36 euros, and maximums of 92 lines and 89100 euros.

	Min	1st quartile	Median	Mean	3rd quartile	Max
Lines	1.000	1.000	2.000	2.317	3.000	92.000

Value	0.07	20.8	36	159.7	101.92	89100.00

Table 3.1, numeric summary of lines and value

In figure 3.4 we can see the makeups of Treatment, store and digital client. We can see that around 80% of clients were classed as silver during their first purchase, 60% made their first purchase in the store, almost all were not digital and around 70% picked it up with their own means.

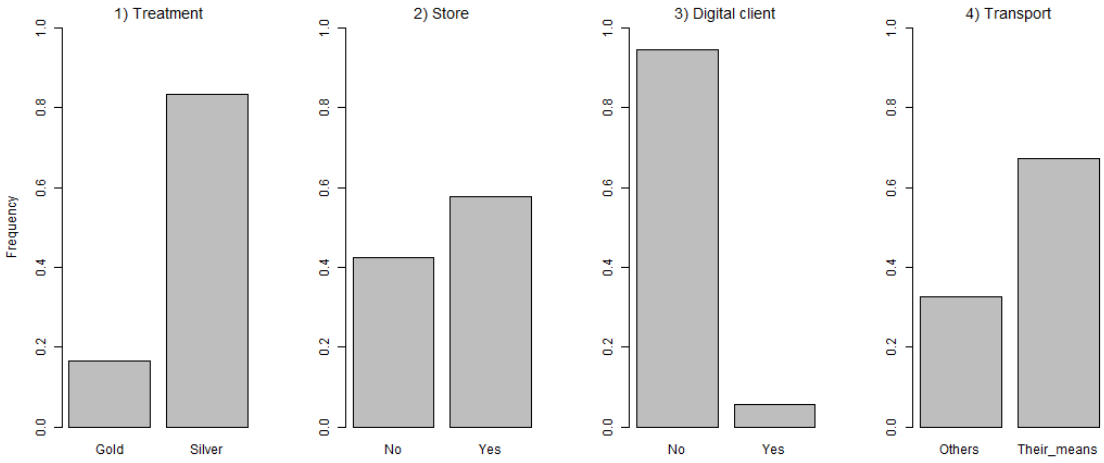


Figure 3.4, bar plots of treatment, store and digital client

We will now move on to the variables related to the clients. In figure 3.5 we can see that MRO makes up around 80% of clients, other is extremely rare and OEM and SI are around 10% to 15%.

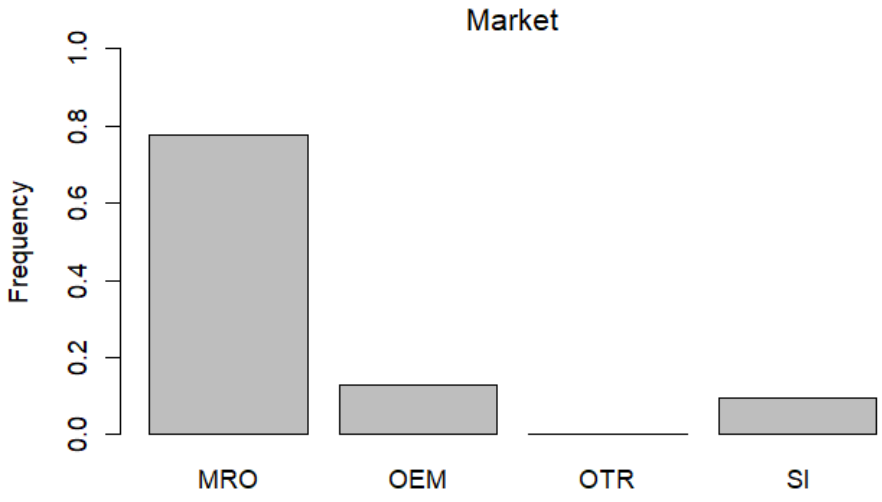


Figure 3.5, bar plots of market

In figure 3.6 we can see that Barcelona has the busiest delegation, followed by Sevilla although at much lower levels, while Tarragona, Pontevedra and Asturias have the lowest amount of clients.

In figure 3.7 we can see how many clients belong to each province, where we grouped all provinces with less than 300 clients in the others category. This way we can see that figures 3.7 and 3.6 are extremely similar, and only Navarra seems to have a large amount of clients despite not having its own delegation.

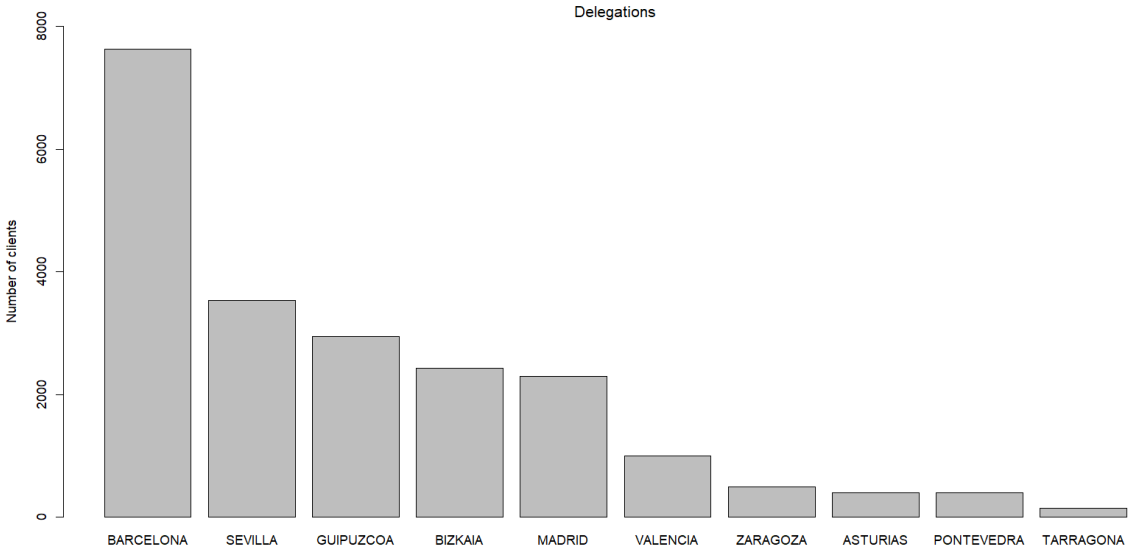


Figure 3.6, bar plot of delegations

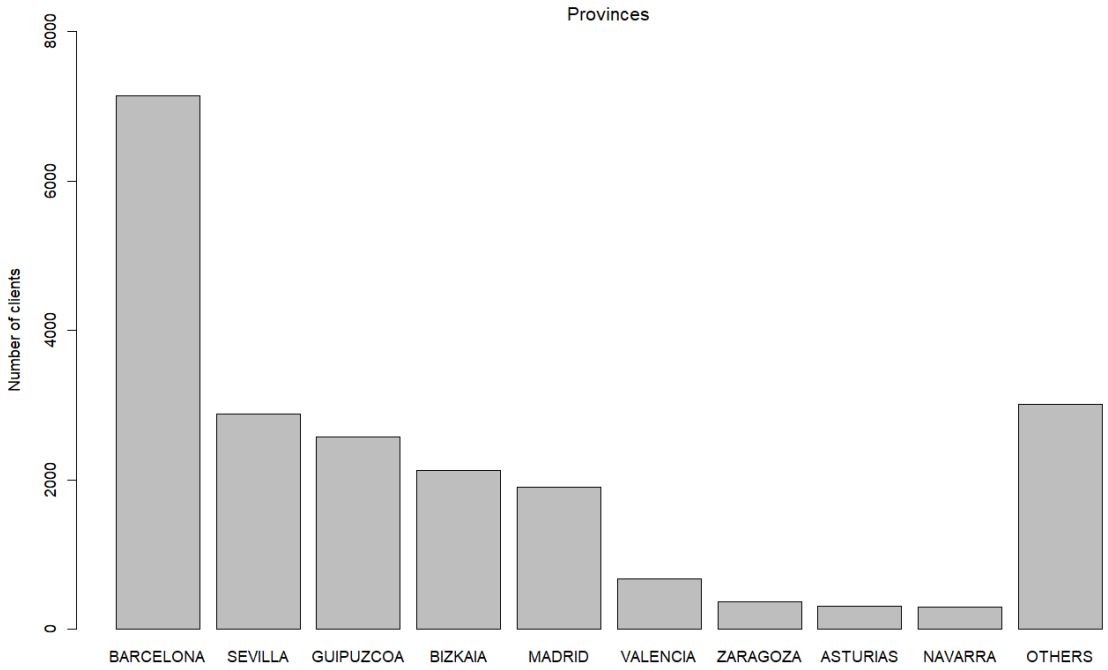


Figure 3.7, bar plot of provinces

Finally, in figure 3.8 we can confirm what we just intuited; the vast majority of clients are assigned to their own province's delegation.

In table 3.2 we can see a numeric summary of the variable distance, which is once again right-skewed, having its median at 0 and the maximum at 9200.

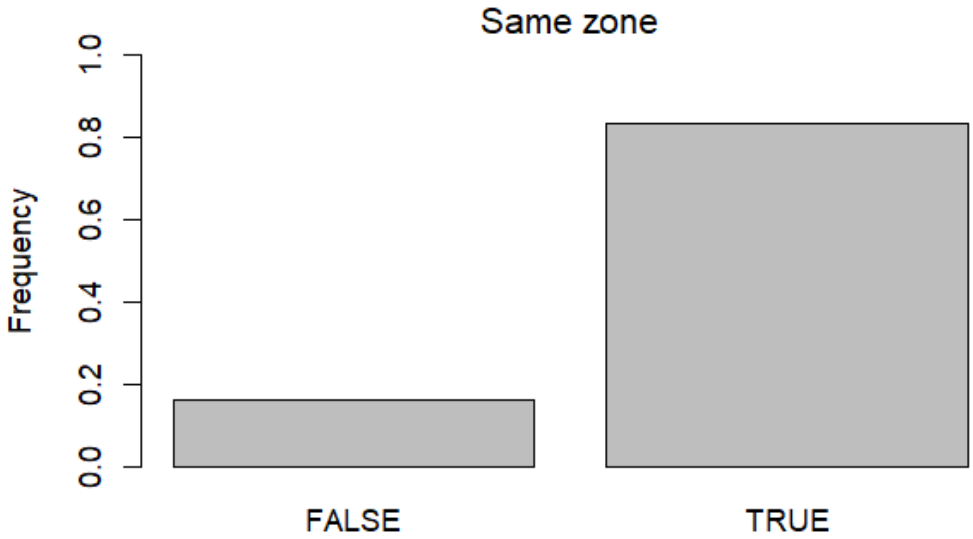


Figure 3.8, bar plot of the variable same zone

Min	1st quartile	Median	Mean	3rd quartile	Max
1.000	1.000	2.000	2.317	3.000	92.000

Table 3.2, numeric summary of the variable distance

Now that we have a better understanding of the data we're working with, we'll move on to the creation of the CLVtools object, with which we'll later fit the model.

3.4. Creating the CLVtools objects

As we said in section 2.3, the Pareto/NBD model is made to model cohorts of clients acquired at a certain time. In our case, given that we have a large amount of data spanning 10 years and customers with very large interpurchase times, as we'll see bellow, we decided to study yearly cohorts, that is, groups of clients acquired in a certain year.

As we mentioned before, from 2010 to 2012 we've got an unusual amount of new clients, because of which we decided to study four different cohorts, clients acquired during 2012, 2013, 2014 and 2015. When creating the object we must specify what fraction of the available time will be used for estimation and which will be reserved for validation. So that these cohort are comparable, and that the latest cohort still has a year of validation data, we will make them all estimate using the year their cohort was acquired and three additional years for estimation.

Having said that, we must remind ourselves of the intended purpose of our model. Wanting to create a statistical tool to classify clients after they are acquired, considering four whole

years of purchase data to create this tool seems dishonest. Because of this we decided to consider two additional cohorts, the clients acquired during 2016 and 2017. For these we will maintain one year of validation data, and the remaining three and two years for estimation, respectively. Their purpose will be to see if using a shorter period of fitting data will have negative consequences in the precision of their predictions and the final classifier.

We then created the CLVtools objects for these six groups of clients, for which we need to use the tickets database and the covariates database we created and explored earlier.

After creating these objects the package allows us to generate some plots to further describe them. Firstly, we can plot the distribution of inter-purchase times measured in weeks for every cohort. In figure 3.9 we can see that for the first cohort these follow a right-skewed and unimodal distribution, with the mode bellow 25 weeks.

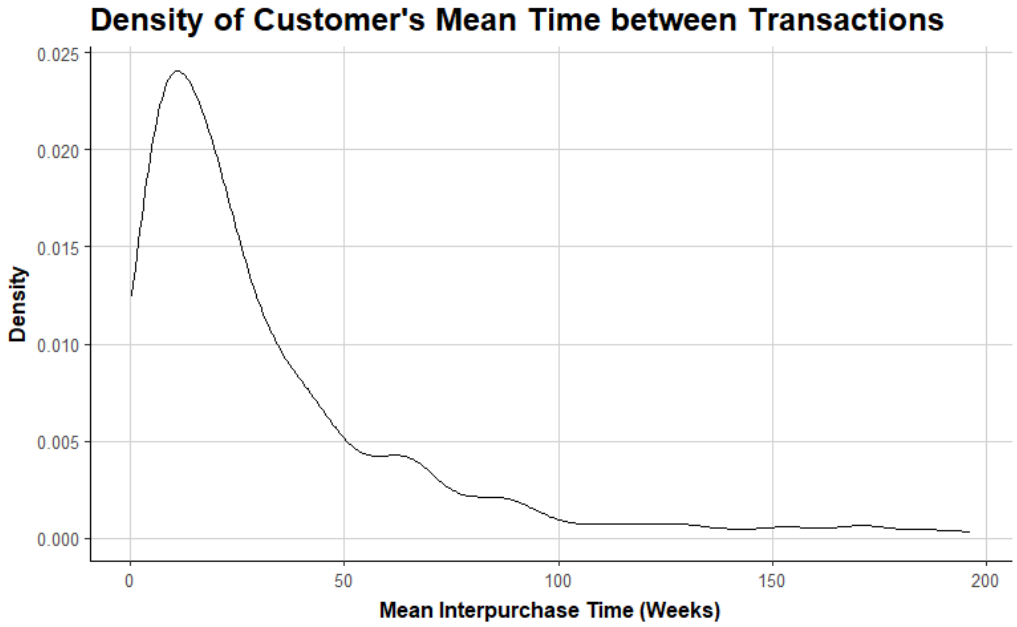


Figure 3.9, Mean time between transactions for the first cohort

All other cohorts give rise to very similar distributions, although the shortening of the fitting period for the last two obviously makes it so the values are smaller. We can see this in figure 3.10 of interpurchase times for the sixth cohort. (The distributions for the rest of the cohorts can be found in figures 7.1 to 7.5 in the annex).

Another type of plots that we are interested by are the bar plots of the amount of repeat transactions. These tell us how many clients never bought again after their first purchase, how many purchased once, etc. In this case, our cohorts seem to be divided in two different groups. As a representative of the first, in figure 3.11 we can see this plot for the first cohort of clients.

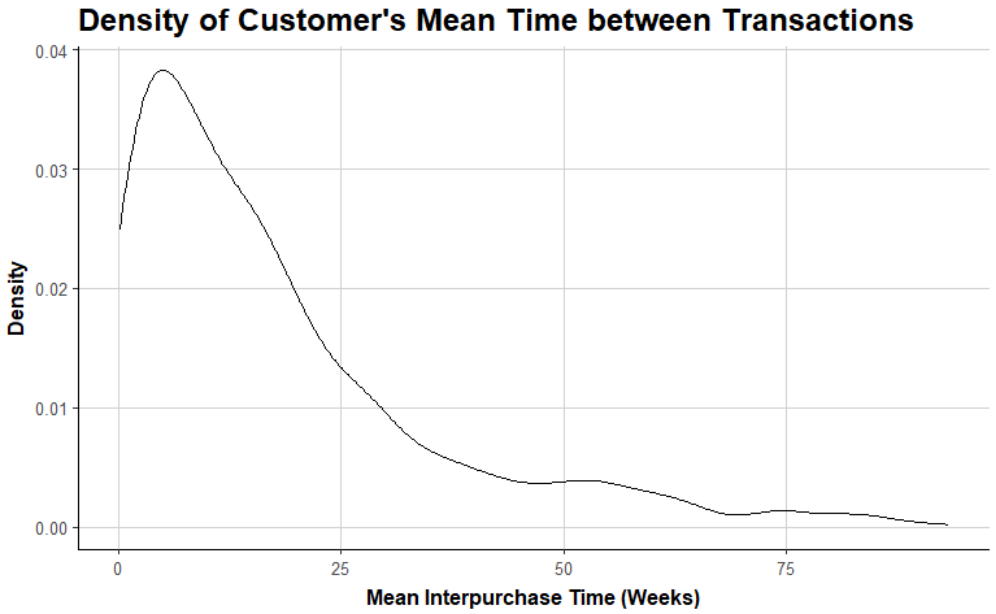


Figure 3.10, Mean time between transactions for the sixth cohort

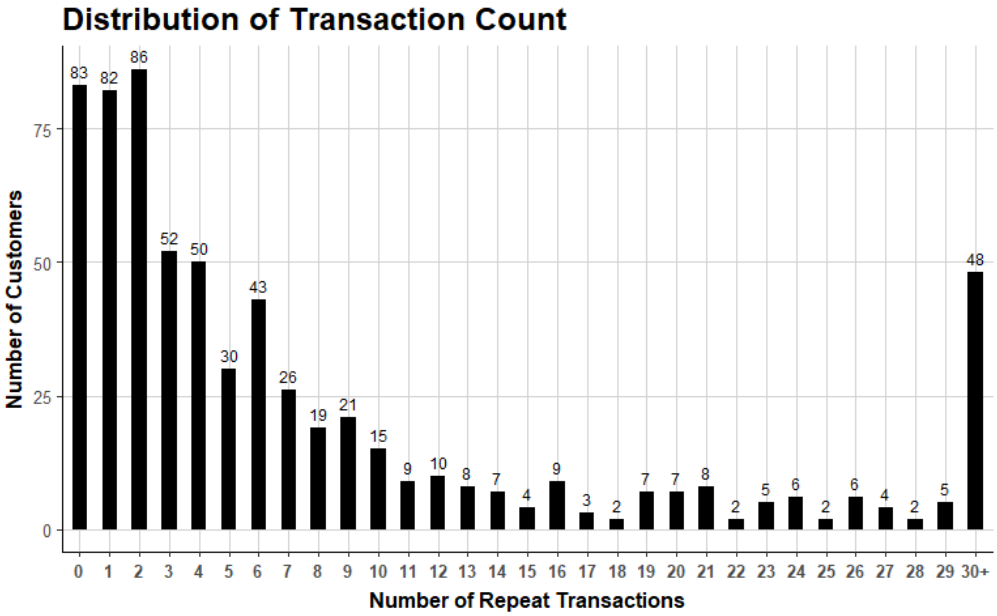


Figure 3.11, repeat transactions bar plot for the first cohort.

For these clients two is the mode amount of repeat transactions and beyond that, the number of customers decreases as the number of repeat transactions increases, although somewhat roughly. This plot is very similar to those of the second and third cohorts (see figures 7.6 and 7.7 in the annex). In contrast, we can look at figure 3.12, where we see that for the sixth cohort zero repeat transactions is the mode, and the decreasing trend is much more observable. The fourth and fifth cohorts exhibit this behavior too, as we can see in figures 7.8 and 7.9 in the annex.

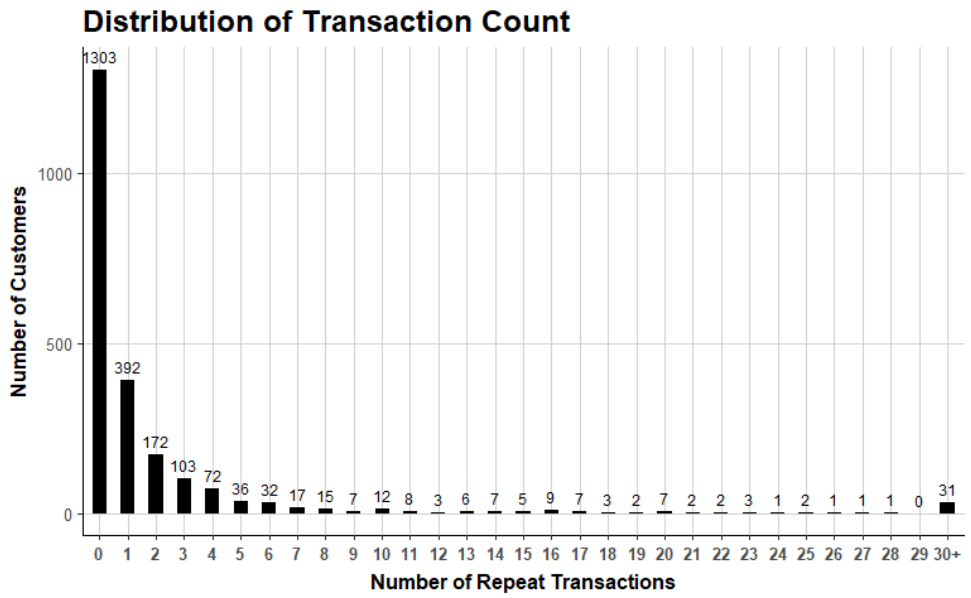


Figure 3.12, repeat transactions for the sixth cohort

Although it is impossible for us to know why the later groups have a much larger amount of clients with no repeat transactions, we can speculate about the composition of the client portfolio and how it was altered by the increase in client acquisition that was experienced from 2012 to 2015. Perhaps from 2015 onwards access to the company's products was greatly increased and as a result, a much larger number of casual clients with no intent of purchasing again started flowing in.

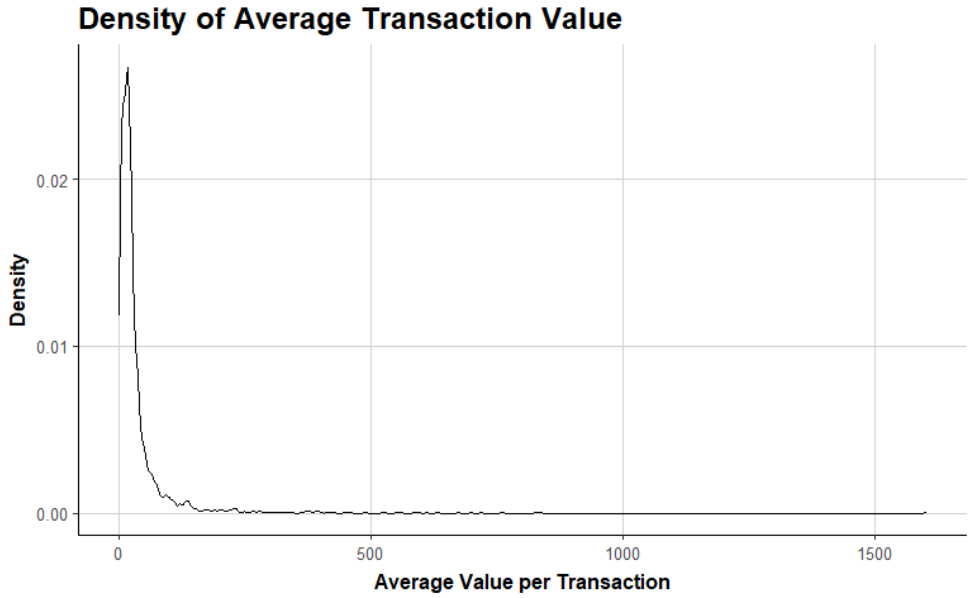


Figure 3.13, mean transaction value for the fourth cohort.

Finally, this package also allows us to plot the mean transaction value of the clients. These follow a distribution very similar to that of the first transaction values described above, as we can see in figure 3.13, and the only difference between the cohorts seems to be how high the

largest value is. The figures for the other cohorts can be found in figures 7.10 to 7.14 in the annex.

3.5. Fitting the Pareto/NBD models

In this section we will fit the Pareto/NBD to the six different cohorts, starting with the simple model and then adding covariates. As we will see, this will be the most challenging part of this thesis, as the numerical properties of the likelihood function for this model are limiting.

3.5.1. Simple models

This becomes apparent the moment we fit the simple models. We do this without giving a specific starting value for any of the coefficients, as we do not have any prior information. After fitting them the estimation for the first three cohorts fails, and instead of the optimal result we obtain an approximation using the Nelder-Mead method, which doesn't validate the Karush–Kuhn–Tucker (KKT) conditions, and NAs in place for the standard deviation of the coefficients and their p-values.

This however doesn't happen for the fourth, fifth and sixth cohorts, which are properly fitted. What the issue might be becomes apparent in table 3.3 of coefficients of the different models.

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6
r	0.5877	0.6088	0.4734	0.29086	0.22710	0.29512
α	8.4059	9.4184	9.3174	6.81391	3.74683	5.90027
s	0.8257	1.3457	1.3929	0.26912	0.19805	0.23873
β	444.1939	604.0913	365.7913	12.84206	5.14648	4.02244

Table 3.3, simple model coefficients for each cohort

Let us recall that β is the rate parameter for the Gamma distribution of the lifetime process while s is its shape parameter, meaning $\frac{s}{\beta}$ is the average dropout rate. This means that the estimated average dropout rates are, for each cohort: 0.0018, 0.002, 0.004, 0.02, 0.04 and 0.05, so for the first three cohorts the average dropout rate is an order of magnitude lower.

We can also see a big difference when it comes to the plots presented in 3.11 and 3.12. If we add the predicted repeat transactions to them we obtain figures 3.14 and 3.15, which are respectively representatives of the first three and the last three cohorts (see figures 7.15 to 7.18) . In the former the amount of zero repeat transactions is radically over-estimated, while this isn't the case in the latter.

It looks as if the Pareto/NBD model only allows to properly model groups of clients whose repeat transaction bar plots are strictly decreasing and have zero as their mode, and that the change in client composition experienced in 2015 that we mentioned before gave rise to a modellable portfolio of customers.

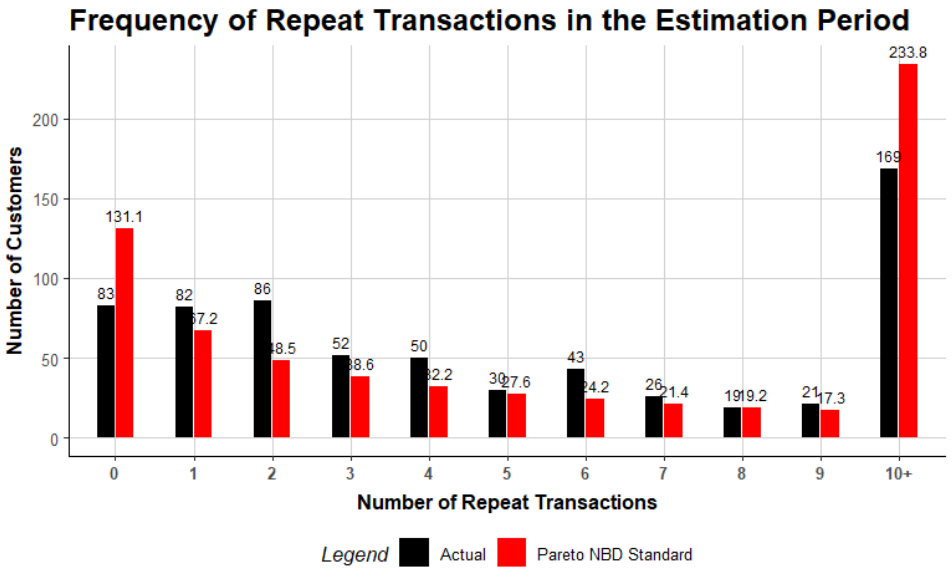


Figure 3.14, observed and predicted repeat transaction counts for the first cohort

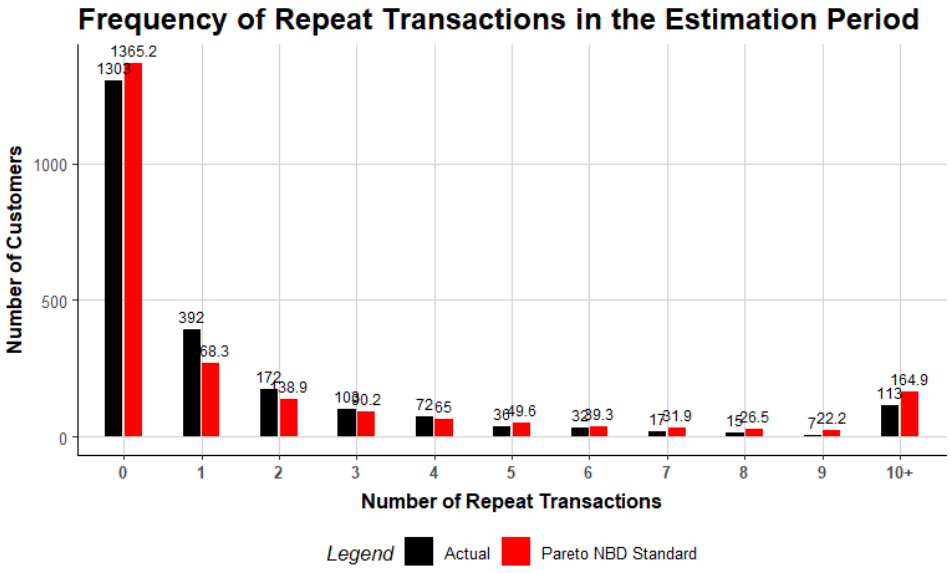


Figure 3.15, observed and predicted repeat transaction counts for the sixth cohort

Although we can't be sure this is the case, since we believe there are sufficient supporting arguments and because properly fitting the model for the first three cohorts is not possible, we'll conclude that those clients don't validate the model assumptions and carry on focusing exclusively on the last three cohorts.

For those three cohorts we can plot the Gamma distributions of the purchase and lifetime coefficients (figures 3.16 and 3.17) and see that there aren't any major differences. We can also calculate the average purchase rates while alive, which are calculated as $\frac{r}{\alpha}$ and take the values 0.04, 0.06 and 0.05 respectively. Both of these and the average dropout rates lead us to think that the three cohorts have similar compositions and that these fitted models are stable.

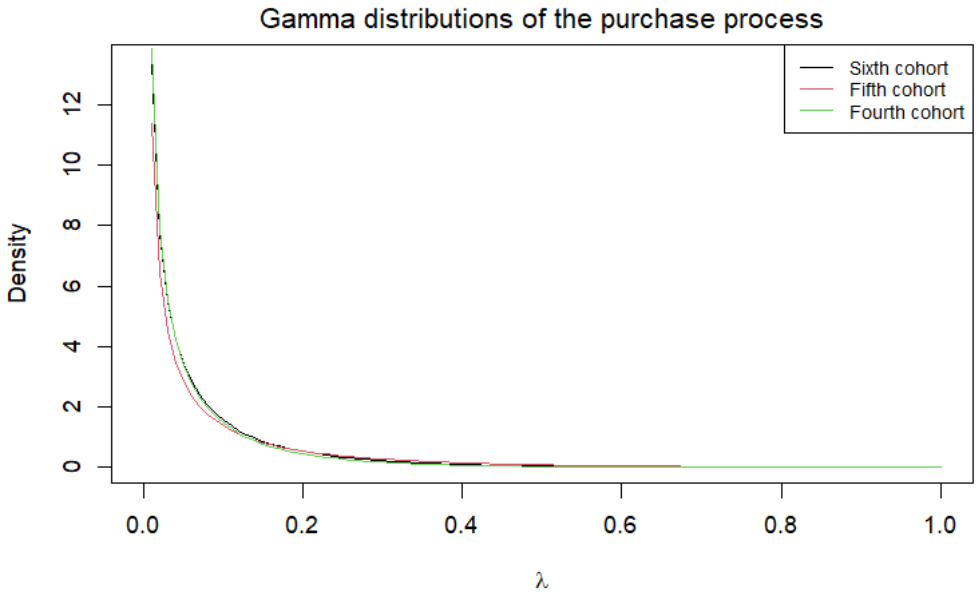


Figure 3.16, Gamma distributions of the purchase process in the last three cohorts

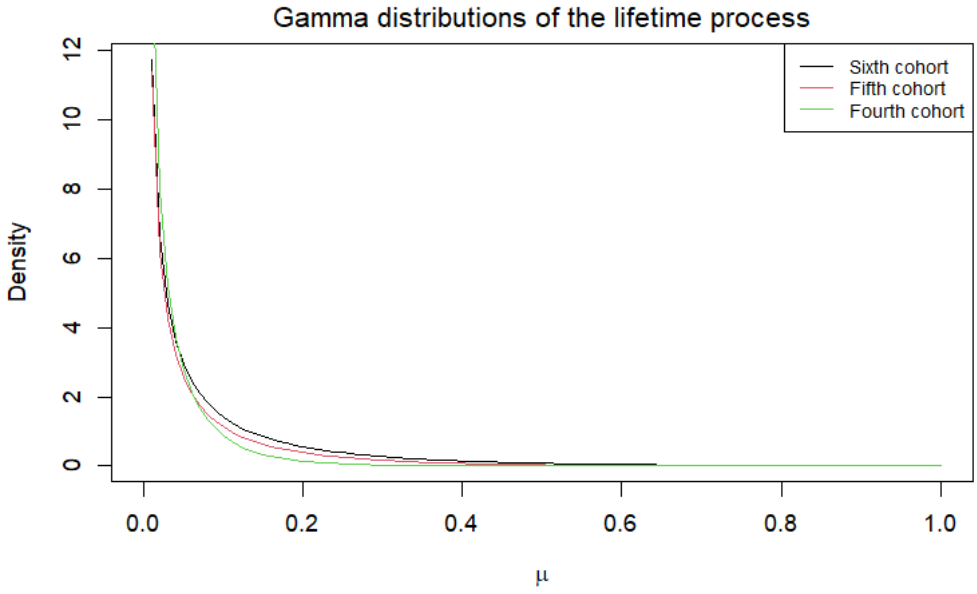


Figure 3.17, Gamma distributions of the lifetime process in the last three cohorts

We can also use the fitted coefficients to calculate the expected amount of purchases in a year (52 weeks) for a randomly chosen customer in each of the groups using the formula we presented before in section 2.3, which are 1.7, 2.3 and 1.7 respectively.

We can also visualize the graph that represents the observed and expected amount of weekly transactions for the fourth cohort. As we can see in figure 3.18, the model seems to properly capture the decreasing amount of transactions through time, both during the fitting and validation periods, although it can't properly represent the high variability from week to week. This result is identical for the fifth and sixth cohorts, see figures 7.19 and 7.20.

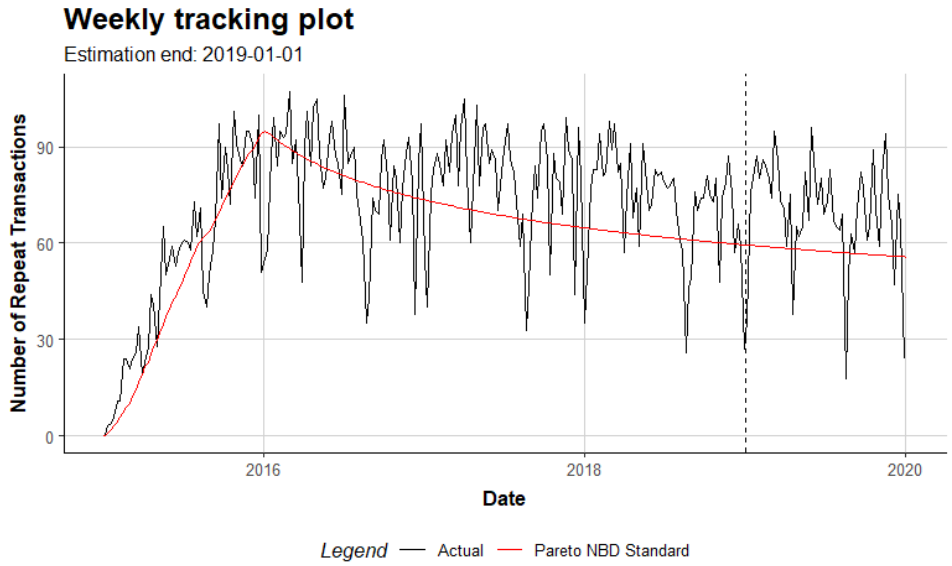


Figure 3.18, expected and observed weekly transactions for the fourth cohort, both during the fitting and validation period, which are delineated by the discontinuous line

3.5.2. Covariate models

We now move on to fitting the models with static covariates. After what we saw in the last section, we'll only work with the last three cohorts. Our initial idea was to adjust a model with every covariate in both the purchase and lifetime processes and then remove all of those that aren't statistically significant using, for example, a stepwise procedure.

We had issues when fitting the models, perhaps because of the large amount of variables. Firstly, we couldn't obtain the optimal result and had to use the Nelder-Mead approximation. Furthermore, if we included the first purchase value, we would get NAs in place for the standard errors and p-values of every coefficient. We imagined it was caused by the variable's large variance and because of this we categorized the variable into its quartiles, which solved the issue for the third and fourth cohorts, but not for the sixth.

After adjusting the complete models we obtain tables where we can see the significance of every variable, both numeric and dummy, in the case of the fourth and fifth cohorts (tables 7.1 and 7.2 in the annex). This however doesn't allow us to evaluate the significance of the

categorical variables, as we only know if there are significant differences between the reference category and the one represented by each dummy.

In order to obtain a p-value associated with each categorical variable we thought about applying the log-likelihood ratio test. This consists in fitting the model with and without each categorical variable and calculating the difference between their respective log-likelihoods. We had to do this manually as the CLVtools object we obtain after fitting the model is of the S4 type and Anova and other existing procedures can only be used for S3 objects.

After calculating it, because we can only work with the Nelder-Mead approximations and not the actual optimal models, we obtain implausible results, since often the smaller model's log-likelihood is bigger than the large one, which invalidates the test.

So we are left with no actual way of reducing the models gradually until all included variables are significant and have to change our approach.

We decided to instead observe the possibility of including a single variable that could offer interpretable results and a greater predictive capacity. Again, we were restrained by the computability of this model. Observing the tables 7.1 and 7.2, we first considered including the variable Market, but when doing so two of the models didn't properly fit and didn't validate the KKT condition. This also happened with the variables distance and number of lines. Because none of these gave manageable results and we didn't want to include covariates in only some of the cohorts, as we wouldn't be able to compare them, we decided to abandon the idea of including any covariates in our models.

3.6. Fitting the Gamma/Gamma models

As we've already fit the Pareto/NBD models, all that we have left is to adjust those relating to the average value of the purchases of every client. Like we explained, we will do so using the Gamma/Gamma spending model.

This tool is also a Bayesian model we can fit using the CLVtools package, although unlike before it can't take covariates, neither static nor dynamic. Therefore, all we need to adjust it is the tickets database.

The models are properly fitted, again without specifying starting parameters, and we obtain the coefficients in table 3.4.

	Cohort 4	Cohort 5	Cohort 6
p	0.60386	0.49689	0.56826
q	1.69334	1.49482	1.54781

γ	58.44245	62.59187	58.19041

Table 3.4, fitted coefficients of the Gamma/Gamma spending models for cohorts 4, 5 and 6

Where, we recall, p is the shape parameter of the Gamma distribution of the spending process and q and γ are the shape and scale parameters of the Gamma distribution that accounts for customer heterogeneity, respectively.

We can see that the three groups of clients give rise to very similar values and distributions. In figure 3.19 we can see the expected and observed distribution of the average values for the fourth cohort, which is very similar to that of the other two cohorts (see figures 7.21 and 7.22), the only difference being how high the highest actual average transaction value is in that cohort.

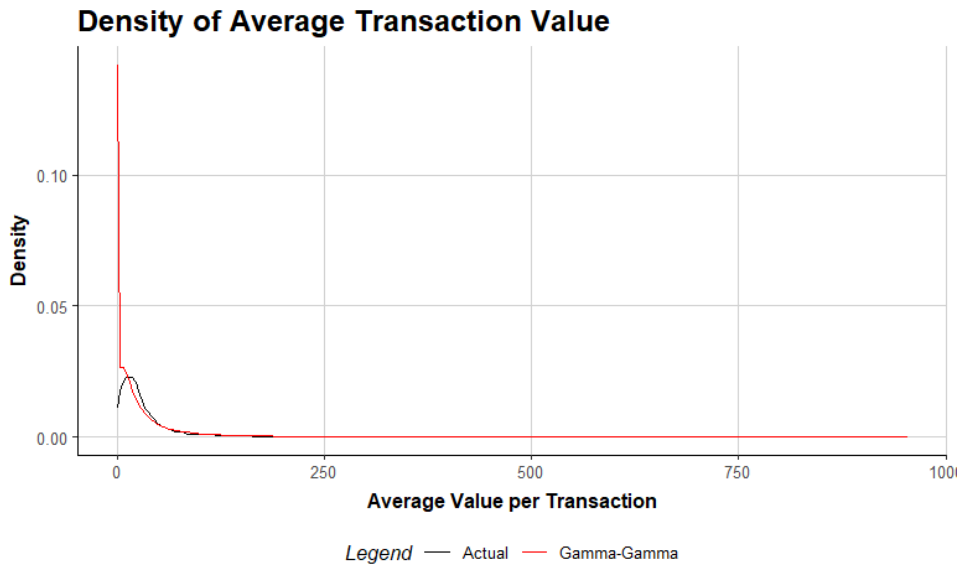


Figure 3.19, expected and observed density distributions of the average transaction value in the fourth cohort.

In the three groups of clients, the model overestimates the amount of values under the actual mode, as it appears it only takes a decreasing shape, although after the mode it seems to fit over the actual distribution well. This means that our average value prediction will be very conservative, as in, most of the errors it will make will probably be due to underestimating, and this will probably affect our final classifier.

Now that we've fitted all the models we need we'll first validate how well these do predicting quantity and value of purchases and then build our good client classifier.

3.7. Model validation

We will now move on to evaluating the validity of our models' predictions. If we use the function "predict()" on the fitted Pareto/NBD model without specifying new data we obtain a

dataframe that includes the model's predictions for our clients on the period of data we reserved for model validation. The function will also automatically fit the Gamma/Gamma model we manually fitted before in order to predict transaction values. The dataframe includes, for every client, the following variables:

- Customer Id.
- First and last date of the validation period (2019-01-02 and 2019-12-31 in our case).
- Length of the validation period measured in our granularity of interest (52 weeks in our case).
- Estimated probability of being alive at the end of the estimation period.
- CET: Conditional expected transitions, that is, number of expected transactions during the validation period.
- DERT: Discounted expected residual transactions, that is, the expected total amount of transactions for the residual lifetime of the customer after the end of the estimation period (how many times they will purchase again).
- Actual x: Actual number of transactions made during the validation period.
- Predicted mean spending: Predicted mean spending per transaction.
- Actual total spending: Actual amount of money spent during the validation period.

We will evaluate how well the Pareto/NBD model predicts the amount of transactions per client during the validation period first, and then how well it predicts the total spent amount, which depends on both the Pareto/NBD and the Gamma/Gamma models.

If we plot the variables CET and Actual x for every cohort we obtain figures 3.20, 3.21 and 3.22, where we can see that there seems to be a solid linear relation between them, although the variability of the data points appears to slightly increase in the later cohorts, which leads us to believe that the shorter adjusting periods might decrease the accuracy of the predictions.

For these two variables we can also calculate the mean square error (MSE) of the predictions, which is defined as $\mathrm{sqrt}\left(\frac{\sum_{i=1}^{n}(e_i-o_i)^2}{n}\right)$, where n is the amount of customers and e_i and o_i are the expected and observed values for the i^{th} customer. We obtain the values 6.081, 9.022 and 10.623 respectively. We can't compare them as they are absolute measures of error, and the variables take slightly different values in the three cohorts, but given the values the range of the variables, they do tell us that the predictions are reasonable in the three cohorts.

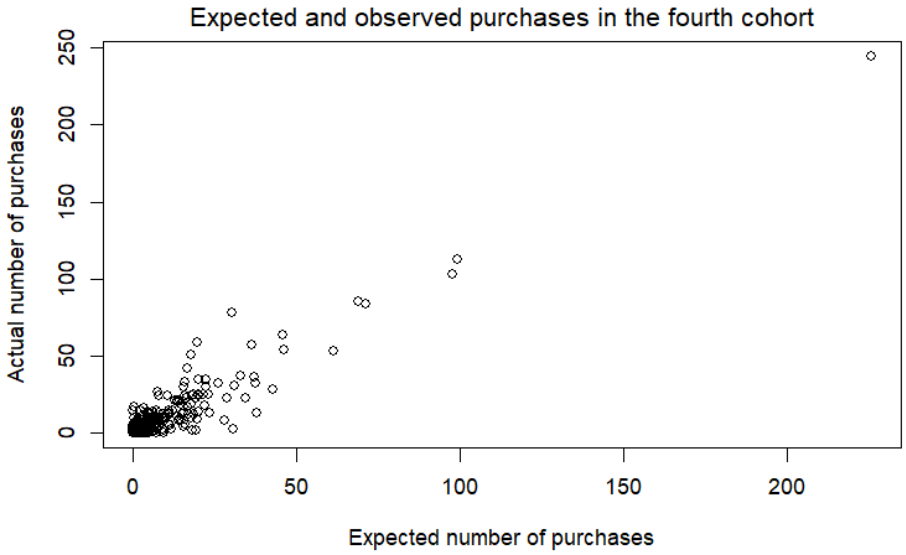


Figure 3.20, predicted and observed amount of transactions per client in the validation period for the fourth cohort.

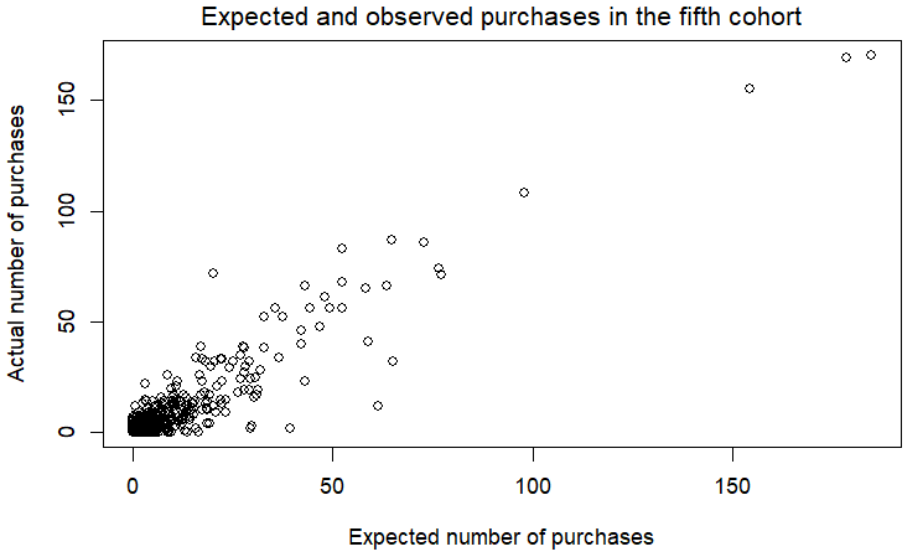


Figure 3.21, predicted and observed amount of transactions per client in the validation period for the fifth cohort.

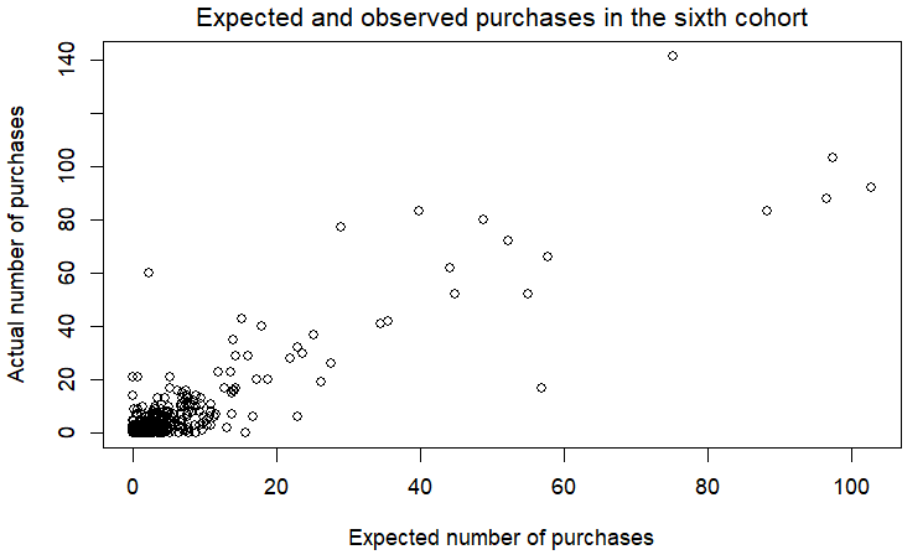


Figure 3.22, predicted and observed amount of transactions per client in the validation period for the sixth cohort.

If we plot the variables actual total spending and the product of CET and predicted mean spending, which is the model's prediction for the customer's total spending during the validation period (2019), we obtain figures 3.22, 3.23 and 3.24. In those graphs we can see that there is still a clear linear relation between the two variables, although variability has greatly increased.

We also calculated the RMSE for every cohort and obtained 1421.838, 523.951 and 768.364. Again, we mustn't compare these values, but they do again tell us that the predictions aren't entirely unreasonable.

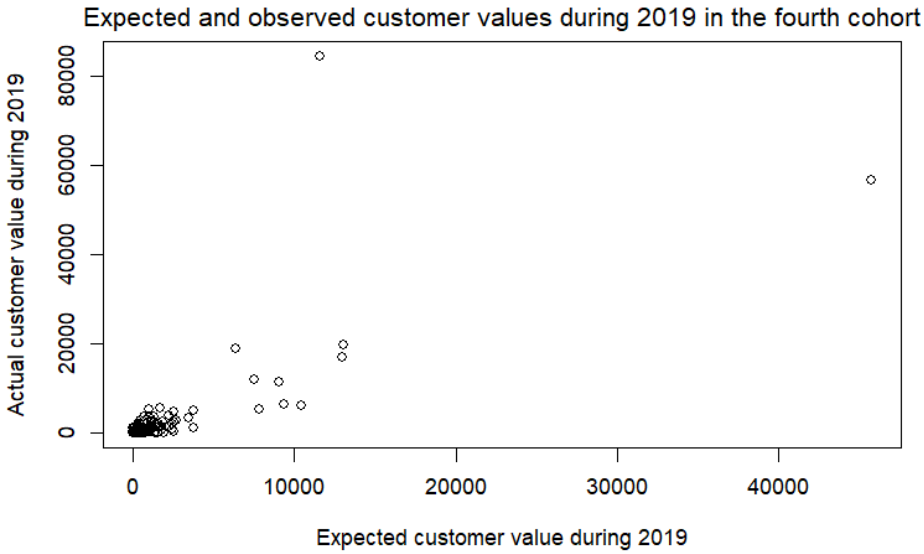


Figure 3.22, predicted and observed customer value per client in the validation period for the fourth cohort.

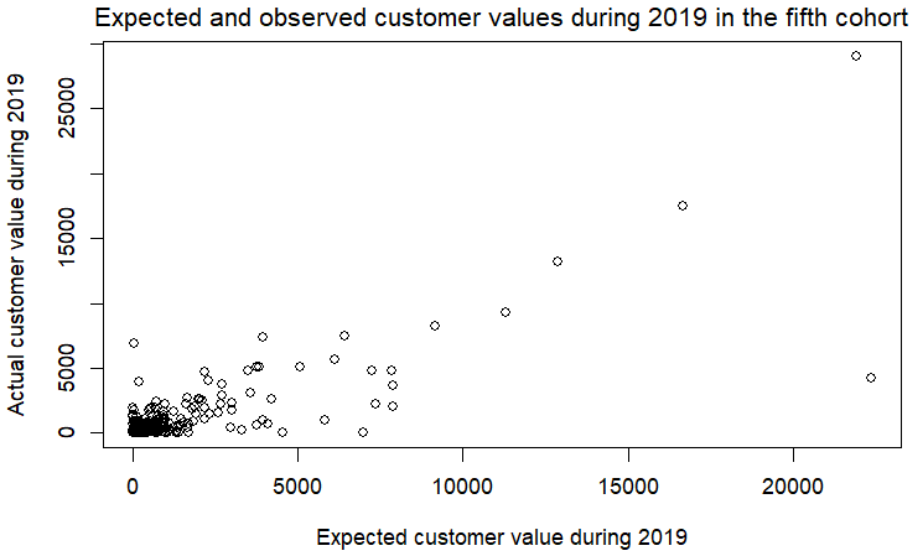


Figure 3.23, predicted and observed customer value per client in the validation period for the fifth cohort.

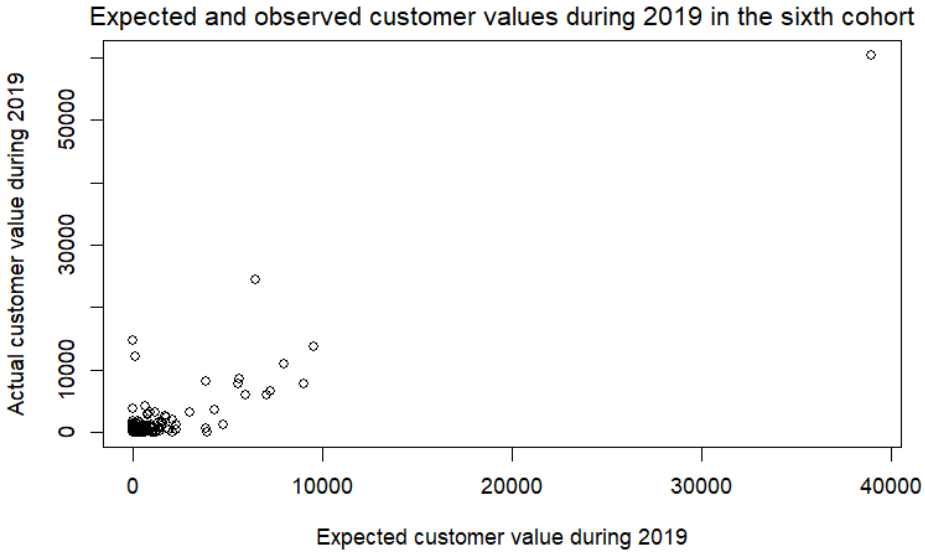


Figure 3.24, predicted and observed customer value per client in the validation period for the sixth cohort

The last result we are interested in is the probability that the client is dead by the end of the estimation period. If we consider that clients with no purchases during the validation period are actually dead, we can plot the ROC curve of the probability as a predictor of that binary characteristic. In figure 3.25 we can see the ROC curve for every cohort, all of which are very similar.

Using these curves we can also find the optimal cutoff points, which, since we consider sensibility and specificity equally valuable, will be the probabilities corresponding to the closest points in the curves to the upper left corner (coordinates (0,1)). In our case these are 0.475, 0.475 and 0.425, respectively, which are fairly similar.

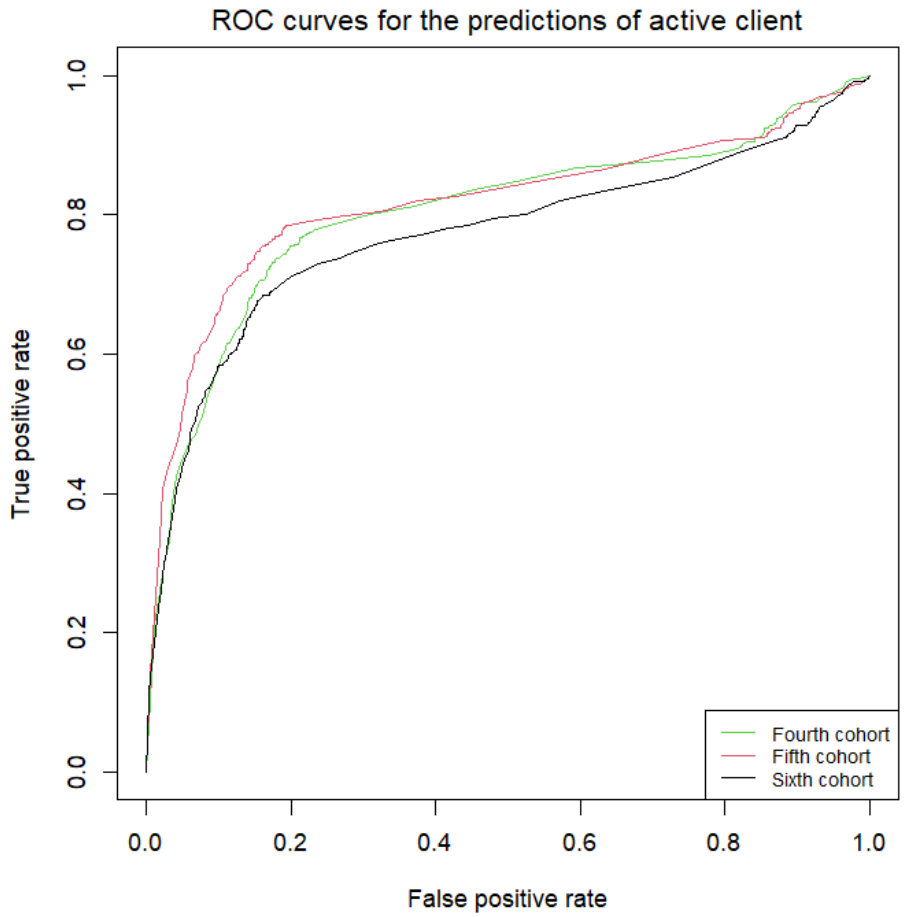


Figure 3.25, ROC curves for the alive client predictor in each cohort.

If we take the optimal cutoff point for every cohort, the active client predictors constructed with these cutoffs will have the following characteristics, as seen in table 3.5. These values are very positive and also seem to imply that the model didn't lose predictive power when it comes to this probability in the later cohorts, despite having a shorter estimation period.

	Cohort 4	Cohort 5	Cohort 6
Sensitivity	0.7671958	0.783557	0.7121212
Specificity	0.7871607	0.8092253	0.799308

Table 3.5, sensitivity and specificity of the active customer predictor in each cohort

We can therefore attest that these models' predictions are reasonable and now move on to creating the classifier that we set out to put together from the beginning.

4. Creating the classifier

We'll create the classifier using the criteria the company initially gave us. That is, we will take into account the predicted probability that the client is alive and the predicted future spending.

Therefore, we will predict a client will be good the following year if their probability of being alive at the end of the estimation period is higher than that cohort's cutoff point and if the product of their CET and predicted mean spending for the following year is higher than 1000€.

In our case, since for every cohort we have 1 year of validation data, we can consider the clients that spent more than 1000€ during that year to be actually good and use this to see how well our classifier would do. If we do so we obtain the confusion matrices presented in tables 4.1, 4.2 and 4.3.

Total number of customers: 2888		Predicted value	
		Good client	Bad client
Actual value	Good client	37	15
	Bad client	11	2825

Table 4.1, confusion matrix of the final classifier for the fourth cohort

Total number of customers: 2829		Predicted value	
		Good client	Bad client
Actual value	Good client	41	24
	Bad client	23	2741

Table 4.2, confusion matrix of the final classifier for the fifth cohort

Total number of customers: 2262		Predicted value	
		Good client	Bad client
Actual value	Good client	23	19
	Bad client	14	2206

Table 4.3, confusion matrix of the final classifier for the sixth cohort

These confusion matrices allow us to calculate the classifiers' sensitivity and specificity, which we can see in table 4.4.

	Cohort 4	Cohort 5	Cohort 6
Sensitivity	0.7115385	0.6307692	0.547619
Specificity	0.9961213	0.9916787	0.9936937

Table 4.4, sensitivity and specificity for the final classifier in each cohort

In all three cohorts the classifier exhibits very good specificity because of the relatively small amount of false negatives, which means that it detects bad clients properly. The sensitivity, however, steadily decreases with the estimation period length, going form a good value in the fourth cohort, where 71% of good clients were classified as such, to only 54% of good clients being classified as such in the sixth cohort.

This is most likely due to what we saw for the Gamma/Gamma spending model, which was very conservative and underestimated the average spending value of the clients.

Although we would have obviously preferred to also obtain good sensitivity, the high levels of specificity mean this classifier would properly detect bad clients even for cohorts with small amounts of data.

5. Conclusions

Using customer lifetime value models we have been able to create a classifier that detects clients who will remain active in the future and bring in at least a thousand dollars per year. We applied this classifier for multiple customer cohorts and saw that with those for which we have less data sensitivity is reduced but specificity remains extremely high.

Our results aren't directly comparable to those of other students, as we haven't only used data obtained during the first purchase, but rather, the amount and timing of the purchases themselves for the first two, three or four years of relation with the client.

Despite that, we believe this approach is reasonable and applicable by the company. It also has certain advantages, as it allows us to revisit and reclassify old groups of clients with the additional information obtained through the years, which means we can assess the active customer base at all times.

It also has setbacks. Firstly, we weren't able to properly apply this model for customer cohorts whose composition didn't match the model assumptions, and afterwards we weren't able to include the covariates that we believe contain valuable information.

The reasonable performance of our classifier despite the lack of additional information makes us think this approach has a lot of potential and calls for further research. Another avenue through which we would like to see research carry on are the Gamma/Gamma models, as it seems like if we were to obtain a better fit for the average spending values the precision of our predictions would greatly increase.

6. Bibliography

Bachmann P., Kuebler N., Meierer M., Naef J., Oblander E., & Schilter P. (2022). CLVTools: Tools for Customer Lifetime Value Estimation. R package version 0.9.0. URL: https://CRAN.R-project.org/package=CLVTools

Bonilla, J. S. (2021). Análisis de clasificadores para determinar el potencial de clientes nuevos en una empresa industrial. TFM. Barcelona: FME - UPC.

Carbonell, M. (2020). Clústers amb variables mixtes per a la caracterització. TFG. Barcelona: FME - UPC.

Casas, P. (2019). New Customer's Classifier. TFG. Barcelona: ETSEIB – UPC

Fader, P. S., & Hardie, B. G. (2007). Incorporating time-invariant covariates into the Pareto/NBD and BG/NBD models. Retrieved July, 2, 2016.

Fader, P. S., & Hardie, B. G. (2013). The Gamma-Gamma model of monetary value. February, 2, 1-9.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who-are they and what will they do next?. Management science, 33(1), 1-24.

Padilla, N., & Ascarza, E. (2020). Overcoming the Cold Start Problem of CRM using a Probabilistic Machine Learning Approach. Available at SRN: https://ssrn. com/abstract, 2933291.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, URL: https://doi.org/10.21105/joss.01686

Wickham H., François R., Henry L., & Müller K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. URL: https://CRAN.R-project.org/package=dplyr

7. Annexes

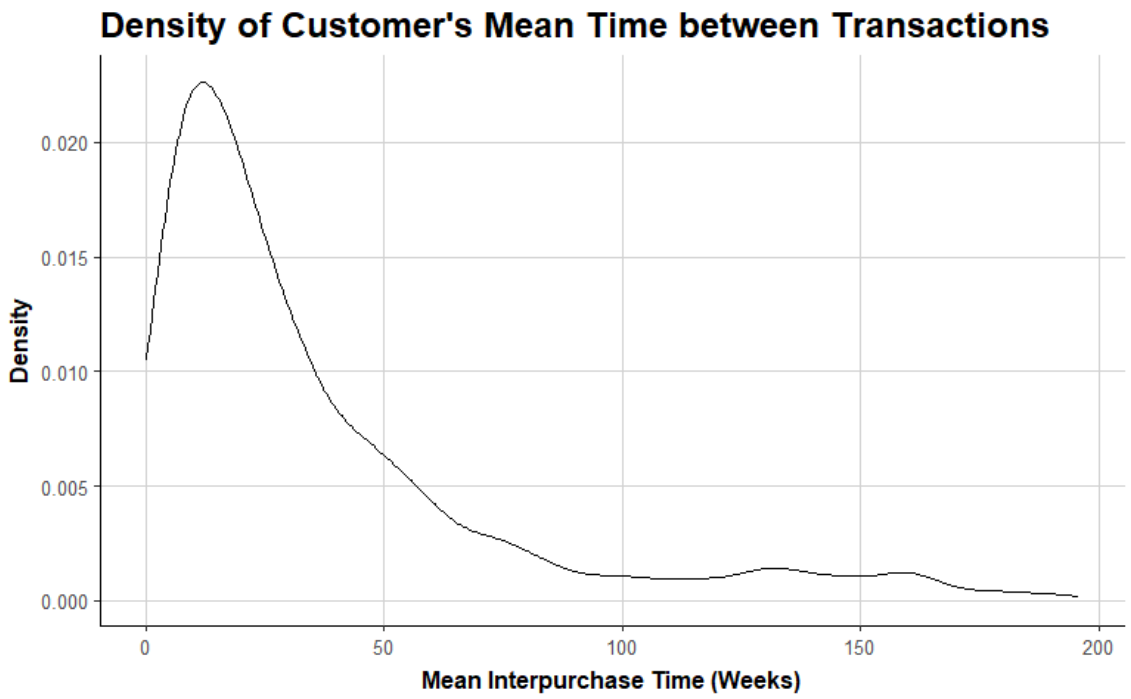


Figure 7.1, Mean time between transactions for the second cohort

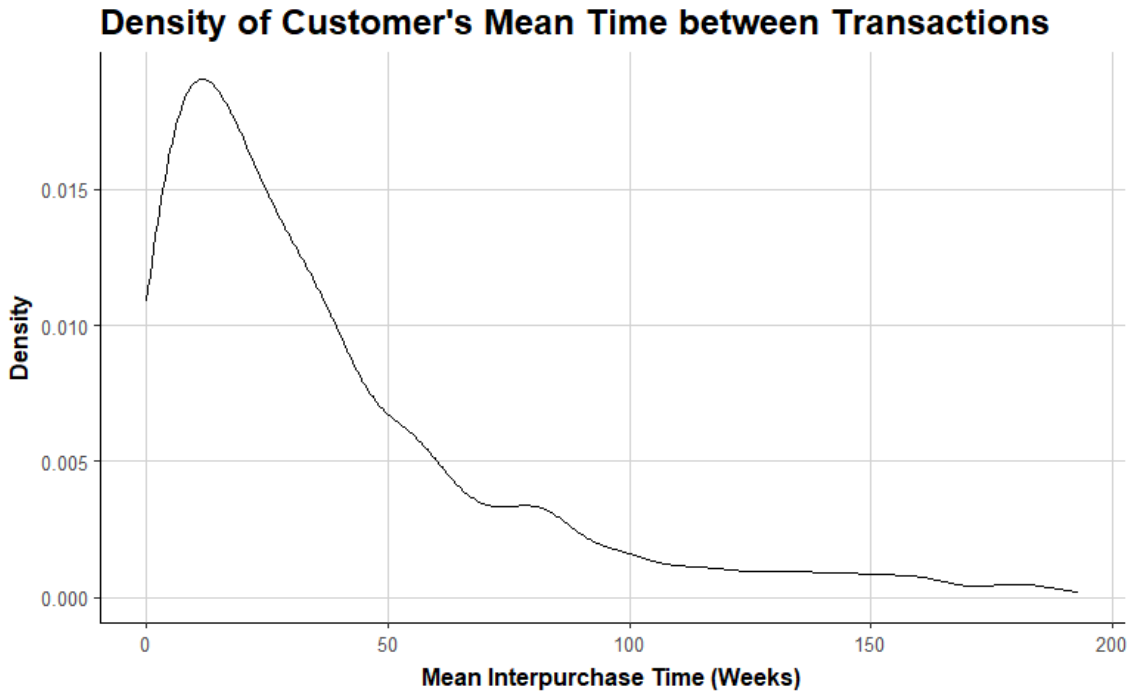


Figure 7.2, Mean time between transactions for the third cohort

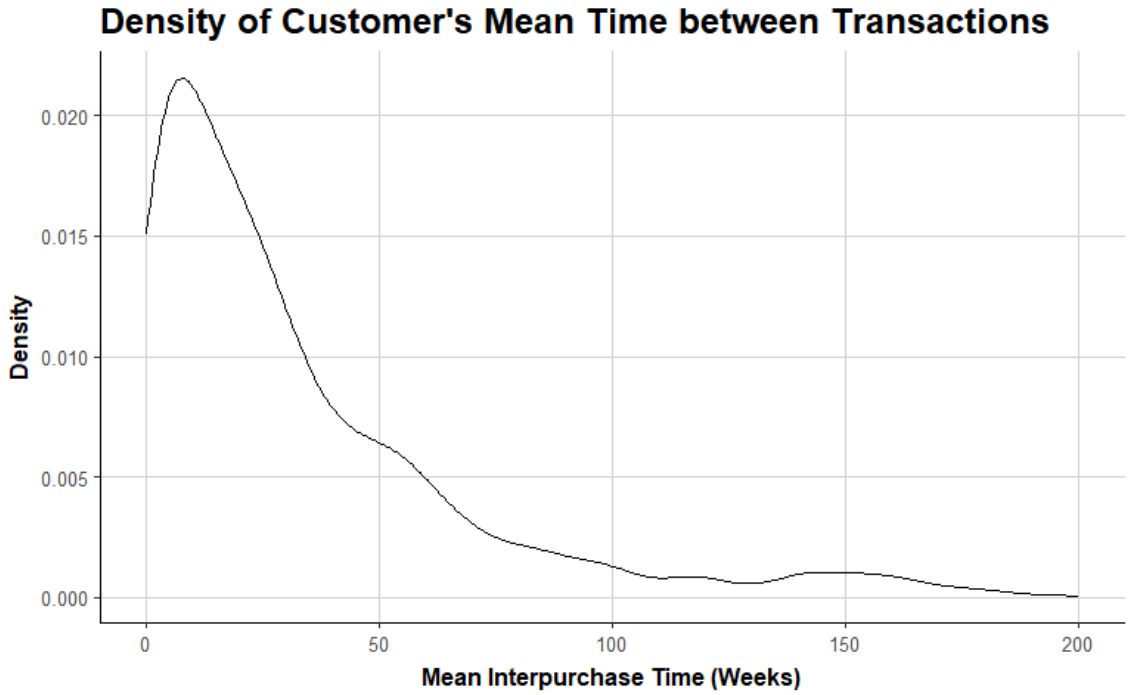


Figure 7.3, Mean time between transactions for the fourth cohort

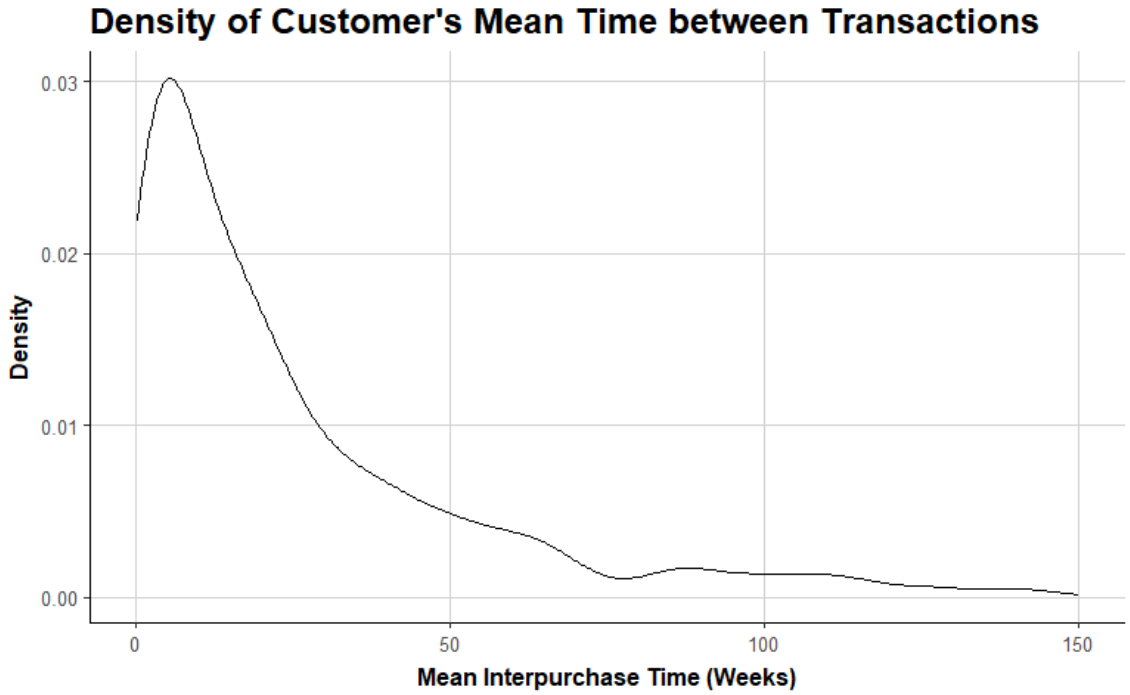


Figure 7.4, Mean time between transactions for the fifth cohort

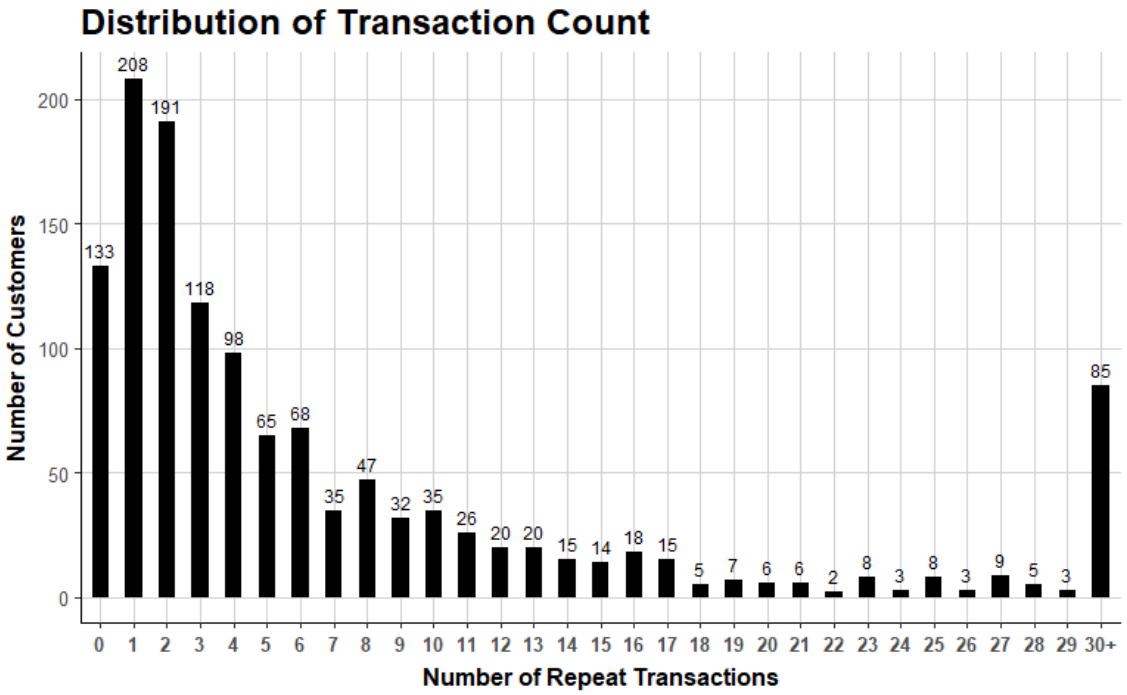


Figure 7.6, repeat transactions bar plot for the second cohort.

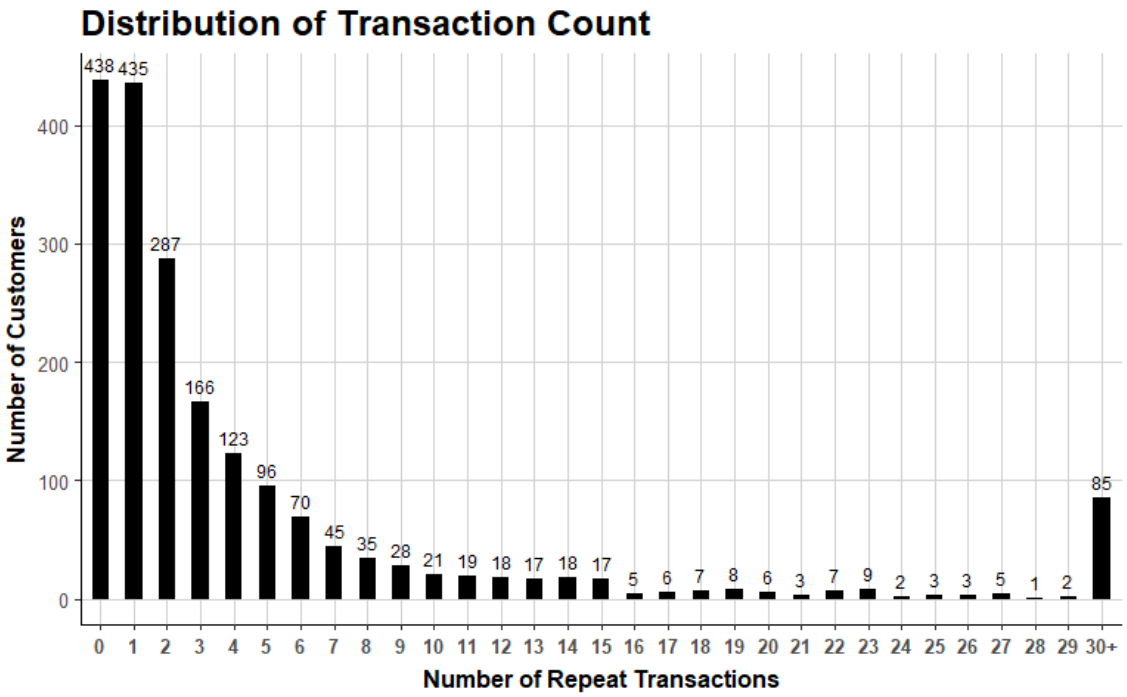


Figure 7.7, repeat transactions bar plot for the third cohort.

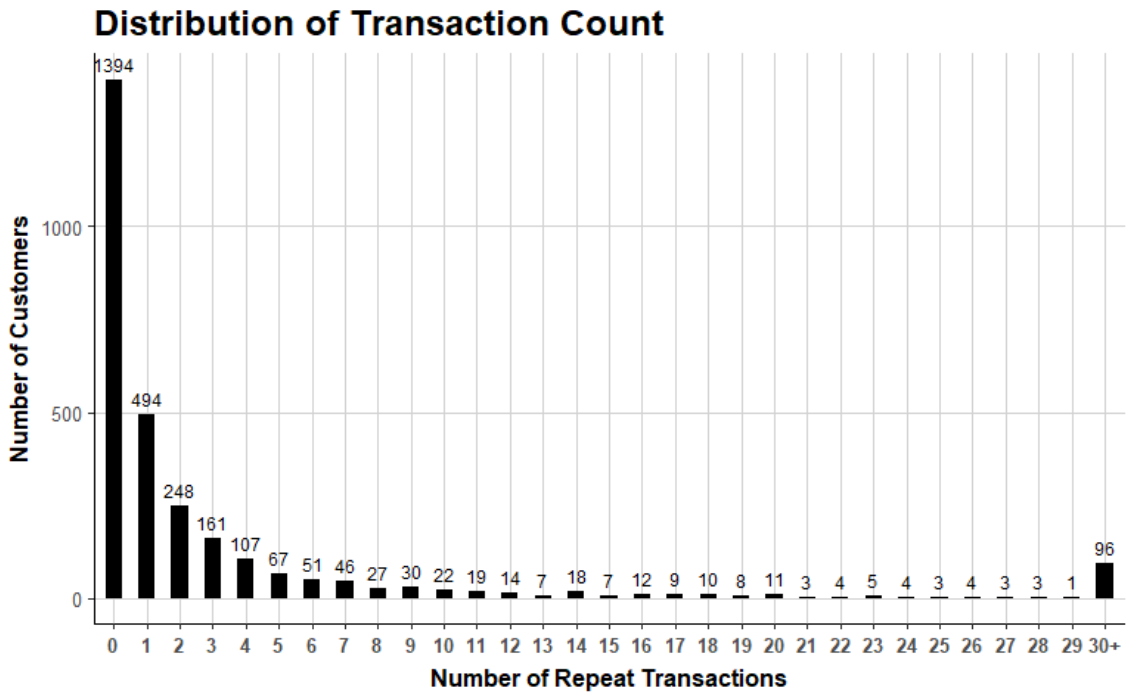


Figure 7.8, repeat transactions bar plot for the fourth cohort.

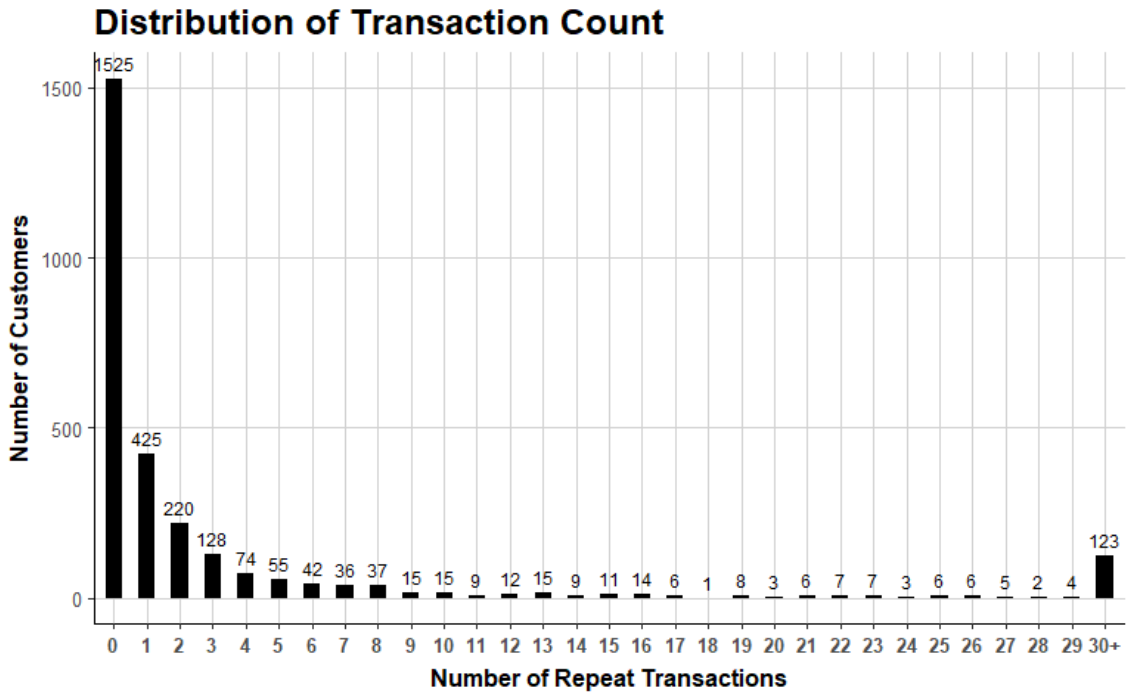


Figure 7.9, repeat transactions bar plot for the fifth cohort.

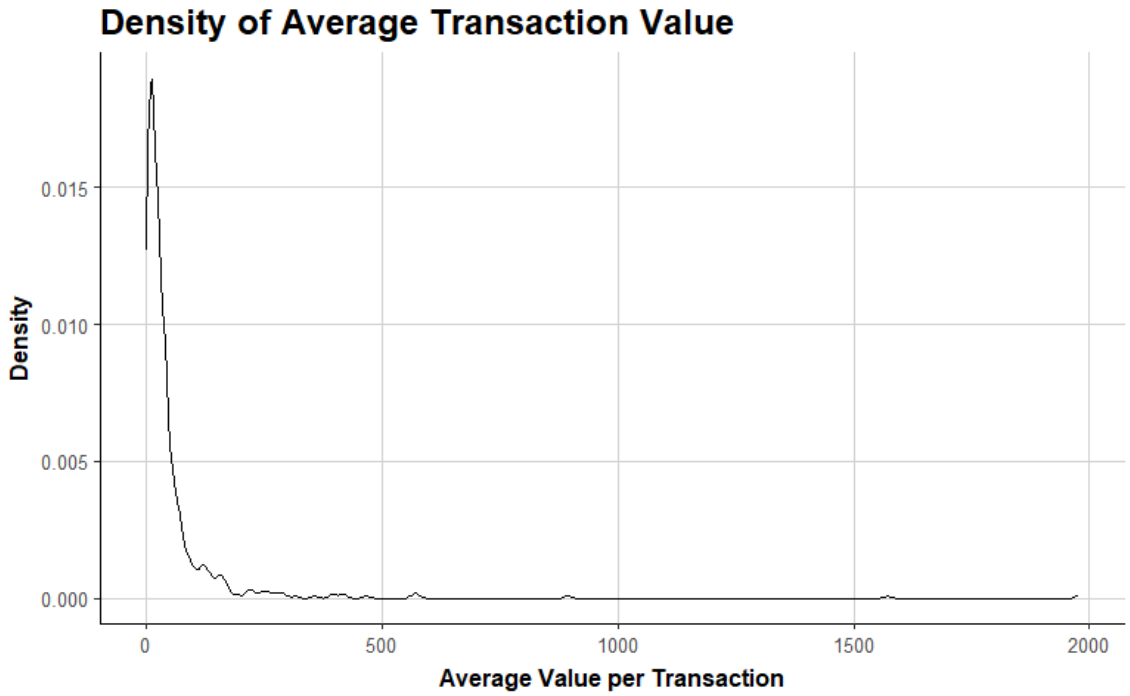
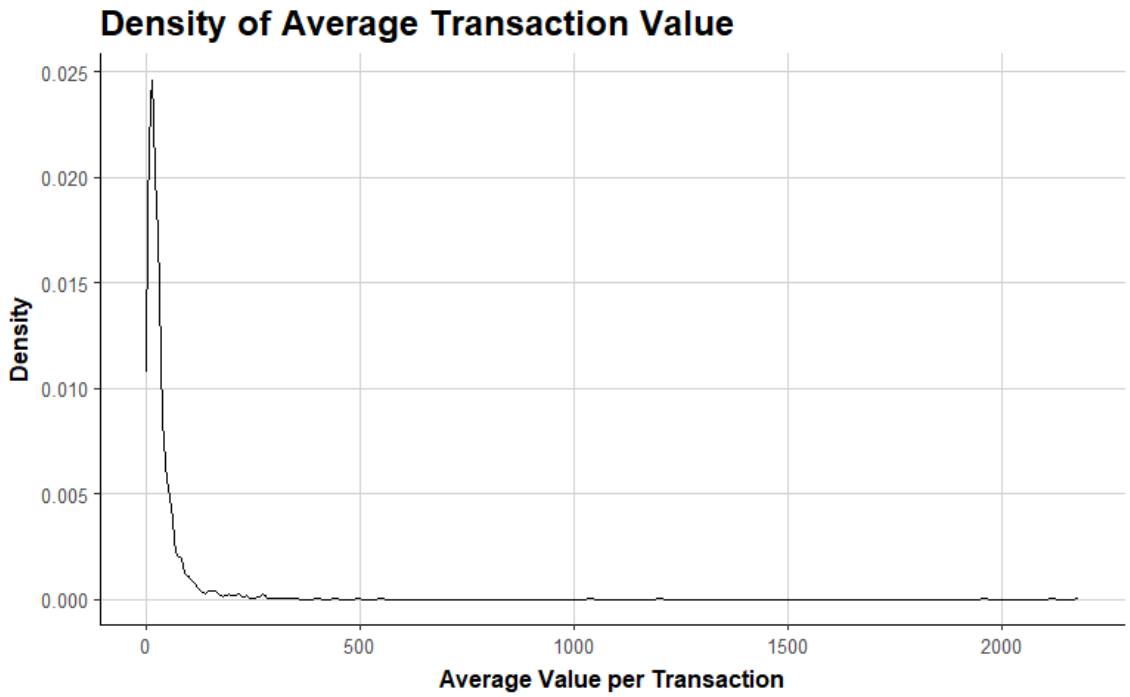


Figure 7.10, mean transaction value for the first cohort.



7.11, mean transaction value for the second cohort.

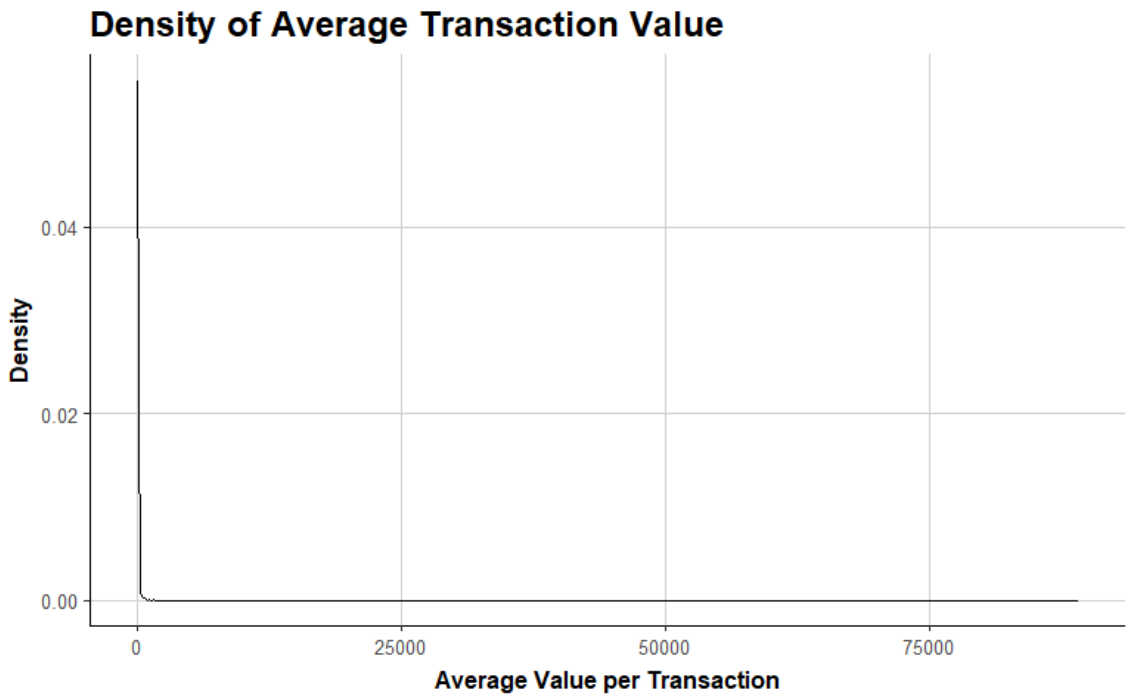


Figure 7.12, mean transaction value for the third cohort.

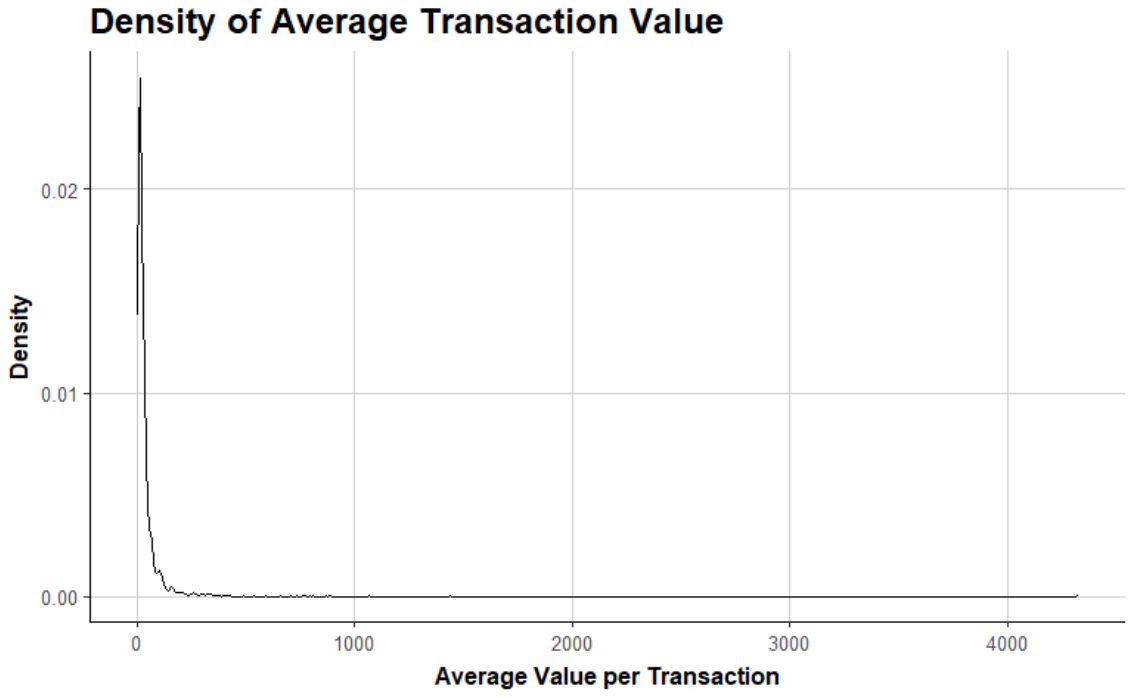


Figure 7.13, mean transaction value for the fifth cohort.

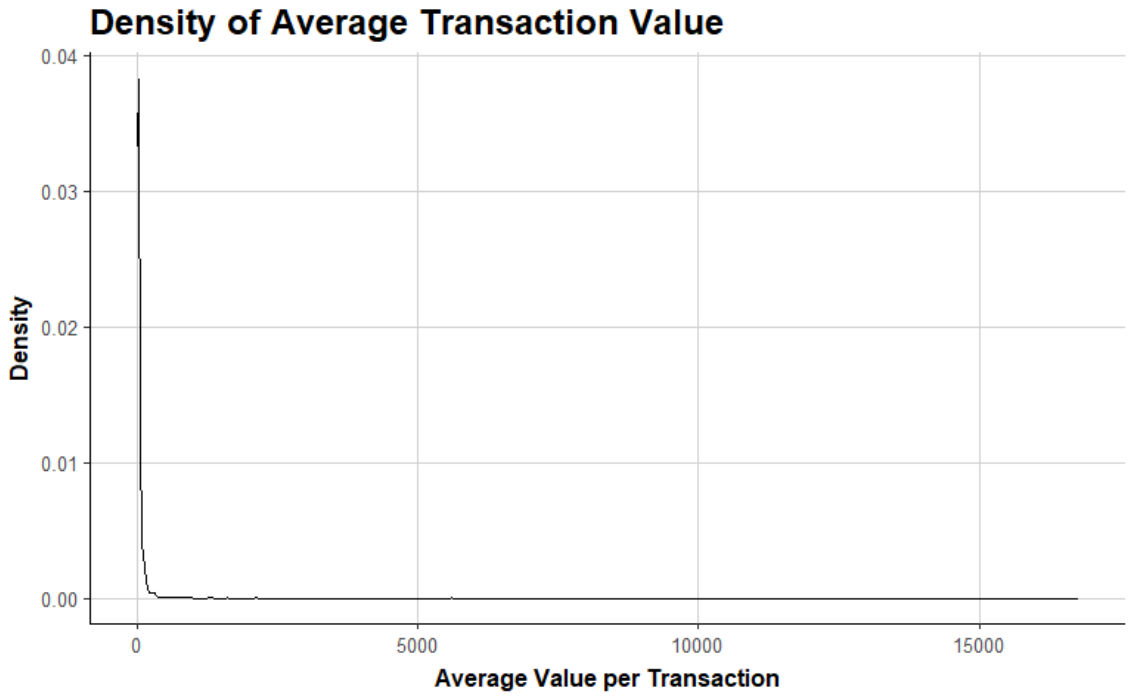


Figure 7.14, mean transaction value for the sixth cohort.

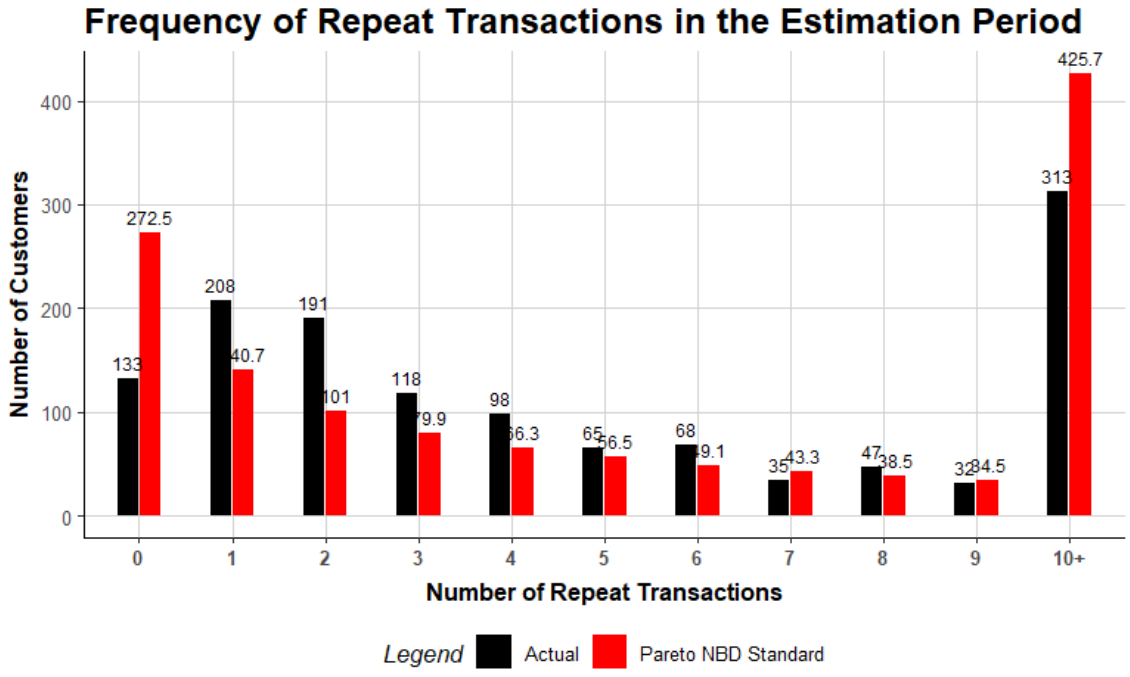


Figure 7.15, observed and predicted repeat transaction counts for the second cohort

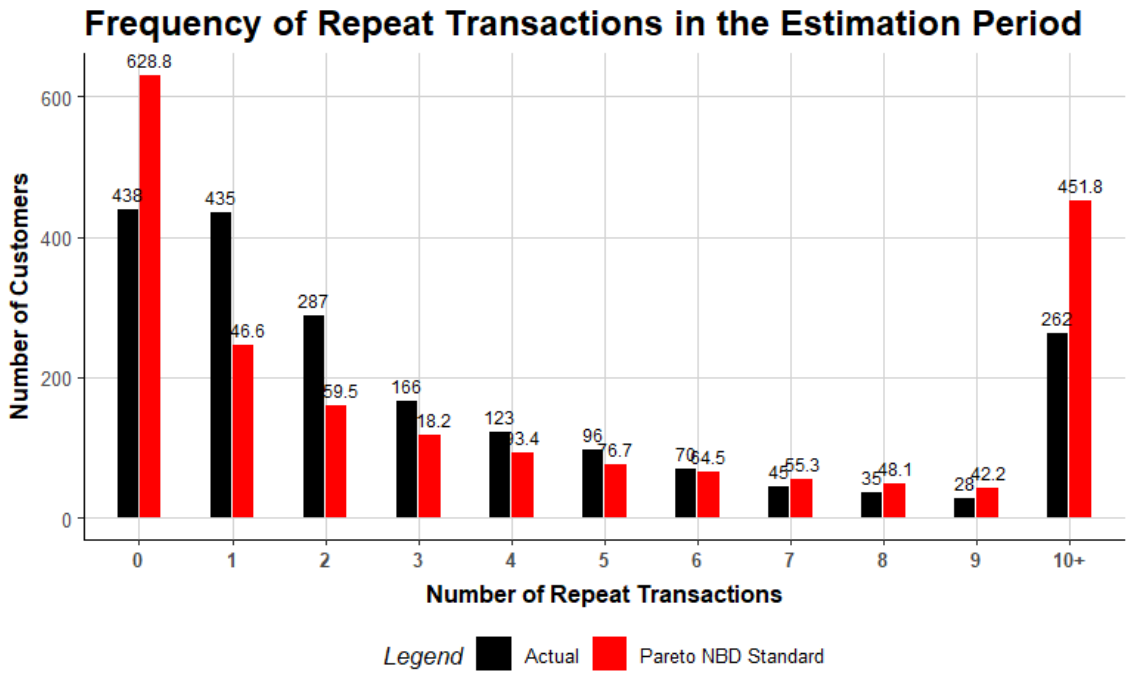


Figure 7.16, observed and predicted repeat transaction counts for the third cohort

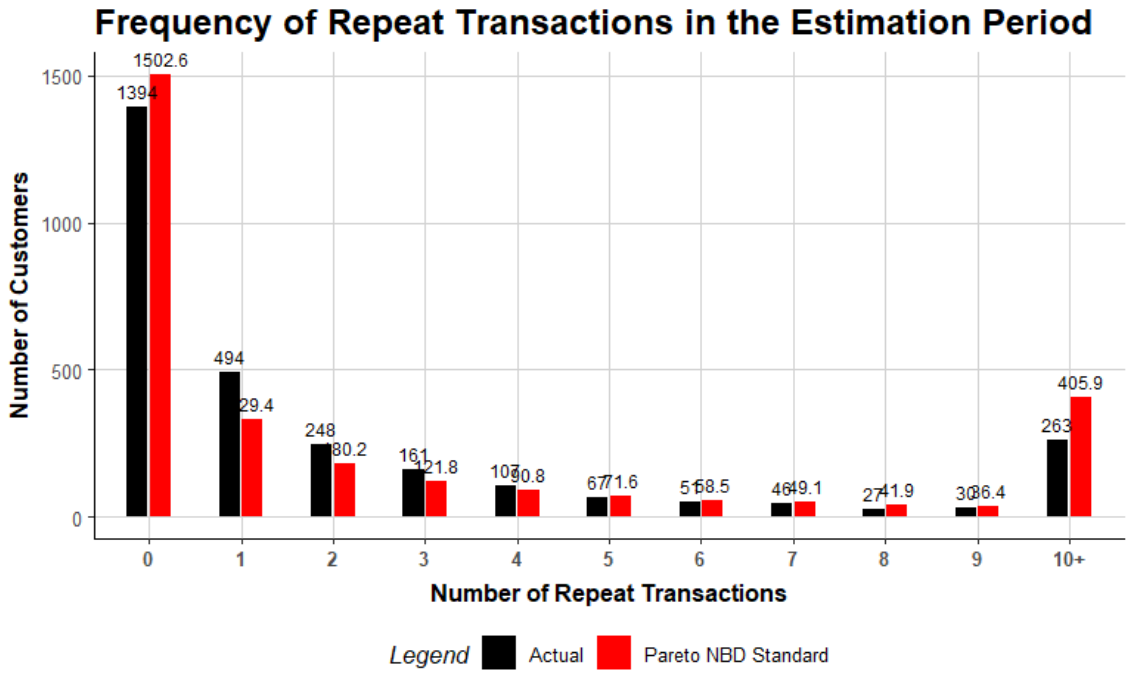


Figure 7.17, observed and predicted repeat transaction counts for the fourth cohort

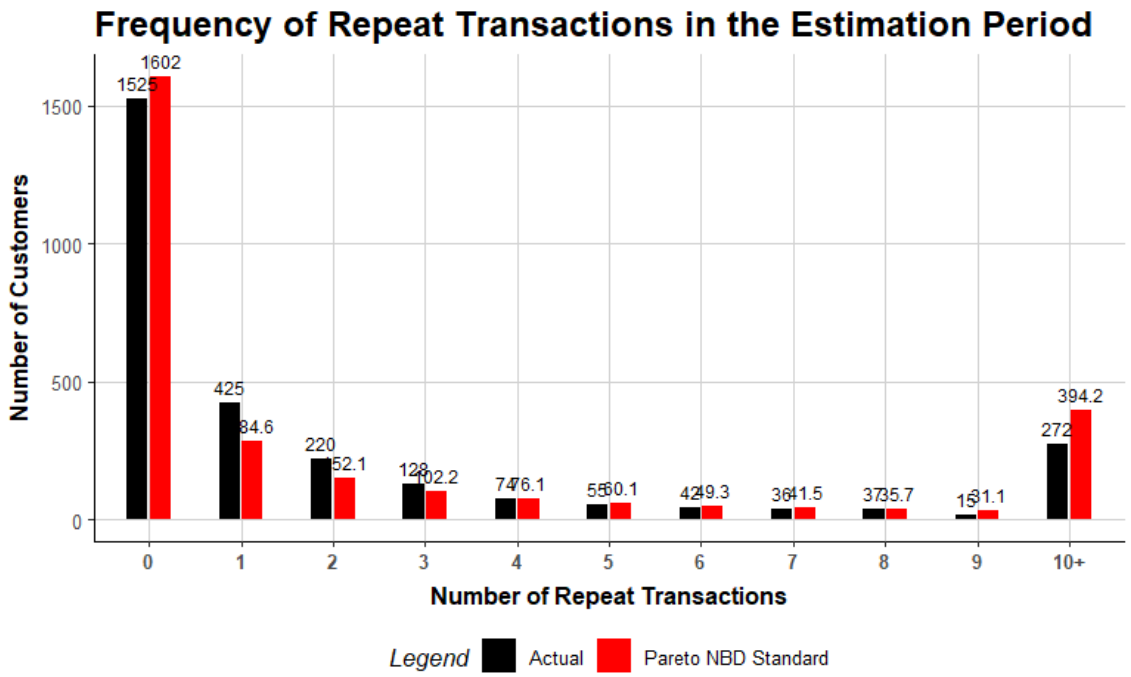


Figure 7.18, observed and predicted repeat transaction counts for the fifth cohort

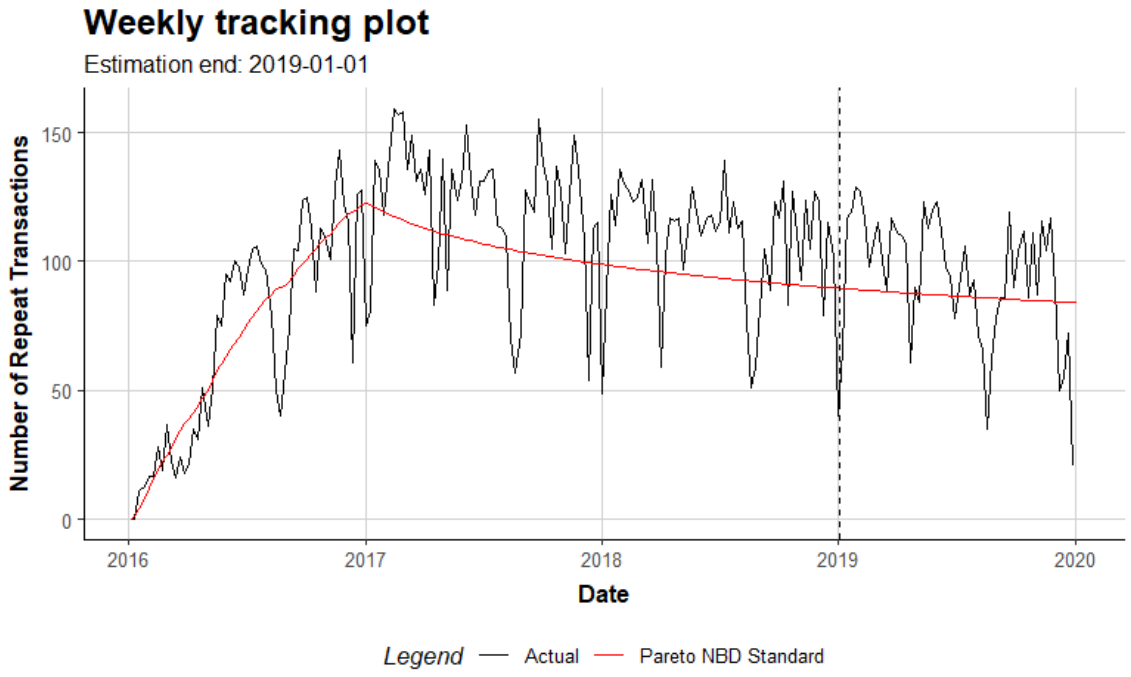


Figure 7.19, expected and observed weekly transactions for the fifth cohort, both during the fitting and validation period, which are delineated by the discontinuous line

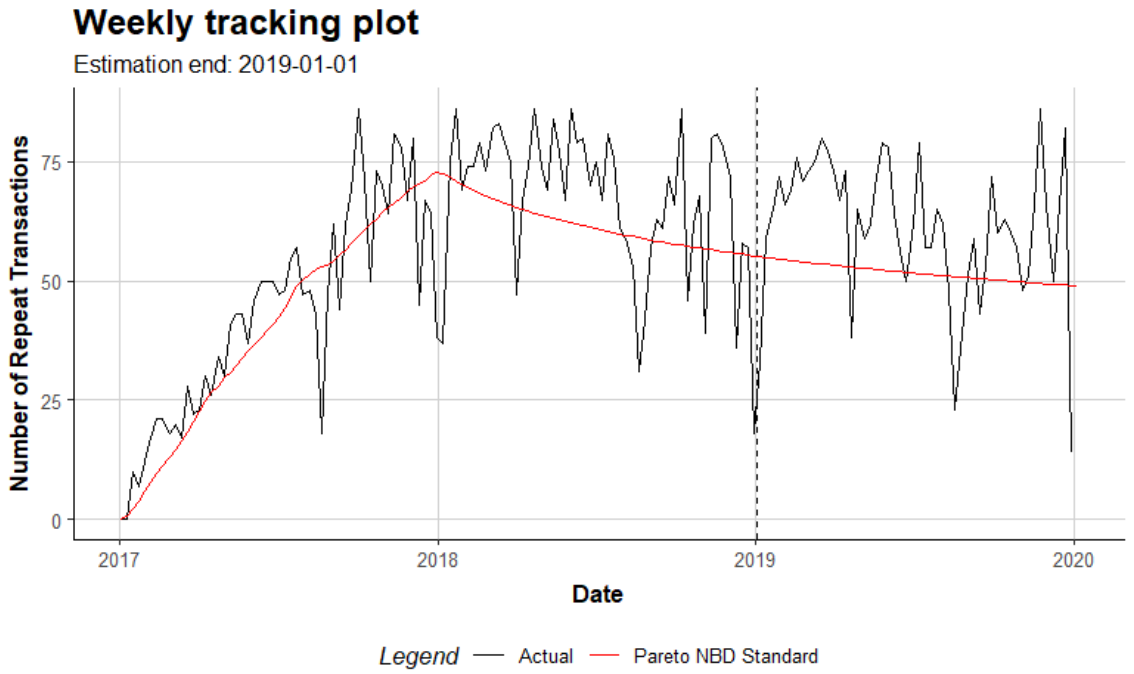


Figure 7.20, expected and observed weekly transactions for the sixth cohort, both during the fitting and validation period, which are delineated by the discontinuous line

	Estimate	Std. Error	z-val	Pr(>|z|)
r	0.9995	0.0113	88.2891	0.0000
alpha	0.9887	0.0254	38.9722	0.0000
s	1.0031	0.0113	88.6908	0.0000
beta	1.0265	0.2153	4.7678	0.0000
life.delegation.BARCELONA	0.1077	0.3342	0.3224	0.7472
life.delegation.BIZKAIA	0.0793	0.4340	0.1828	0.8549
life.delegation.GUIPUZCOA	0.1037	0.3991	0.2598	0.7950
life.delegation.MADRID	0.1019	0.1331	0.7657	0.4439
life.delegation.PONTEVEDRA	0.0935	1.1912	0.0785	0.9374
life.delegation.SEVILLA	0.1075	0.3868	0.2780	0.7810
life.delegation.TARRAGONA	0.1104	1.2979	0.0851	0.9322
life.delegation.VALENCIA	0.0887	0.7465	0.1188	0.9054
life.delegation.ZARAGOZA	0.1311	1.2508	0.1048	0.9165
life.same.zoneTRUE	0.1039	0.5362	0.1937	0.8464
life.fp.transportTheir_means	0.1931	0.3975	0.4859	0.6270
life.distance	0.0666	0.0043	15.3843	0.0000
life.treatmentSilver	-0.1230	0.3609	-0.3409	0.7332
life.marketOEM	0.0877	0.5048	0.1737	0.8621
life.marketOTR	0.0856	0.0002	491.1792	0.0000
life.marketSI	0.0969	0.4135	0.2344	0.8147
life.storeYes	0.1210	0.3401	0.3558	0.7220

	Estimate	Std. Error	z-val	Pr(>|z|)
life.digita_cliYes	0.0903	0.8140	0.1109	0.9117
life.firstpurchase.monthFebruary	0.0981	0.6174	0.1589	0.8737
life.firstpurchase.monthMarch	0.1014	0.4650	0.2182	0.8273
life.firstpurchase.monthApril	0.1286	0.2275	0.5652	0.5720
life.firstpurchase.monthMay	0.1024	0.4085	0.2507	0.8021
life.firstpurchase.monthJune	0.1065	0.5162	0.2064	0.8365
life.firstpurchase.monthJuly	0.1035	0.4427	0.2338	0.8151
life.firstpurchase.monthAugust	0.0965	0.4764	0.2027	0.8394
life.firstpurchase.monthSeptember	0.1081	0.4535	0.2383	0.8117
life.firstpurchase.monthOctober	0.0585	0.4515	0.1295	0.8970
life.firstpurchase.monthNovember	0.1282	0.4372	0.2932	0.7693
life.firstpurchase.monthDecember	0.1140	0.5207	0.2189	0.8267
life.firstpurchase_lines	0.1082	0.0610	1.7721	0.0764
life.categorized_valueFourthQuartile	0.1063	0.2207	0.4817	0.6300
life.categorized_valueThirdQuartile	0.1067	0.2790	0.3823	0.7023
life.categorized_valueSecondQuartile	0.1078	0.2600	0.4148	0.6783
trans.delegation.BARCELONA	0.1369	0.3273	0.4183	0.6757
trans.delegation.BIZKAIA	0.1019	0.4211	0.2420	0.8088
trans.delegation.GUIPUZCOA	0.1014	0.3804	0.2667	0.7897
trans.delegation.MADRID	0.1069	0.2038	0.5247	0.5998
trans.delegation.PONTEVEDRA	0.0879	1.1296	0.0779	0.9379
trans.delegation.SEVILLA	0.1072	0.3734	0.2872	0.7740
trans.delegation.TARRAGONA	0.0813	1.1190	0.0726	0.9421
trans.delegation.VALENCIA	0.1038	0.6693	0.1551	0.8768
trans.delegation.ZARAGOZA	0.0956	1.1842	0.0807	0.9357
trans.same.zoneTRUE	0.0850	0.4842	0.1756	0.8606
trans.fp.transportTheir_means	0.0876	0.3822	0.2291	0.8188
trans.distance	0.0520	0.0040	13.0561	0.0000
trans.treatmentSilver	0.1522	0.3394	0.4484	0.6538
trans.marketOEM	0.1019	0.4158	0.2451	0.8064
trans.marketOTR	0.0889	0.0002	510.1529	0.0000
trans.marketSI	0.1023	0.2608	0.3922	0.6949
trans.storeYes	0.0897	0.3227	0.2778	0.7812
trans.digita_cliYes	0.1069	0.7983	0.1340	0.8934
trans.firstpurchase.monthFebruary	0.0958	0.5010	0.1911	0.8484
trans.firstpurchase.monthMarch	0.1036	0.4365	0.2373	0.8125
trans.firstpurchase.monthApril	0.1025	0.1761	0.5820	0.5606
trans.firstpurchase.monthMay	0.1012	0.3844	0.2633	0.7923
trans.firstpurchase.monthJune	0.1009	0.4972	0.2030	0.8392
trans.firstpurchase.monthJuly	0.1081	0.4112	0.2629	0.7926
trans.firstpurchase.monthAugust	0.1063	0.4548	0.2337	0.8152
trans.firstpurchase.monthSeptember	0.1011	0.4299	0.2353	0.8140
trans.firstpurchase.monthOctober	0.1015	0.4148	0.2446	0.8067
trans.firstpurchase.monthNovember	0.1007	0.4127	0.2440	0.8072
trans.firstpurchase.monthDecember	0.1110	0.4969	0.2235	0.8232

	Estimate	Std. Error	z-val	Pr(>|z|)
trans.firstpurchase_lines	0.1048	0.0596	1.7585	0.0787
trans.categorized_valueFourthQuartile	0.1095	0.3136	0.3492	0.7269
trans.categorized_valueThirdQuartile	0.1060	0.2804	0.3780	0.7055
trans.categorized_valueSecondQuartile	0.1202	0.2649	0.4540	0.6498

Table 7.1, Coefficients of the complete model with all covariates for the fourth cohort, where life indicates the variable is in the lifetime process and trans indicates the variable is in the transaction process

	Estimate	Std. Error	z-val	Pr(>|z|)
r	0.9727	0.0189	51.4593	0.0000
alpha	1.0234	0.0599	17.0784	0.0000
s	0.9892	0.0155	63.6442	0.0000
beta	1.0074	0.0170	59.1352	0.0000
life.delegation.BARCELONA	0.1216	0.1891	0.6434	0.5199
life.delegation.BIZKAIA	0.1728	0.3487	0.4956	0.6202
life.delegation.GUIPUZCOA	0.0934	0.3486	0.2678	0.7889
life.delegation.MADRID	0.1114	0.3485	0.3196	0.7493
life.delegation.PONTEVEDRA	0.0629	0.9459	0.0665	0.9470
life.delegation.SEVILLA	0.1050	0.3310	0.3171	0.7511
life.delegation.TARRAGONA	0.1079	1.2478	0.0865	0.9311
life.delegation.VALENCIA	0.0997	0.4735	0.2106	0.8332
life.delegation.ZARAGOZA	0.1077	0.7343	0.1467	0.8834
life.same.zoneTRUE	0.0960	0.3045	0.3154	0.7525
life.fp.transportTheir_means	0.1106	0.1503	0.7361	0.4617
life.distance	0.0062	0.0030	2.0726	0.0382
life.treatmentSilver	0.1726	0.2609	0.6615	0.5083
life.marketOEM	0.1044	0.3309	0.3155	0.7524
life.marketOTR	0.1223	0.0019	64.1255	0.0000
life.marketSI	0.0944	0.3941	0.2396	0.8106
life.storeYes	0.0969	0.1938	0.5002	0.6169
life.digita_cliYes	0.1067	0.6045	0.1765	0.8599
life.firstpurchase.monthFebruary	0.0941	0.3647	0.2579	0.7965
life.firstpurchase.monthMarch	0.0960	0.4077	0.2354	0.8139
life.firstpurchase.monthApril	0.0538	0.3659	0.1471	0.8831
life.firstpurchase.monthMay	0.1109	0.3491	0.3178	0.7506
life.firstpurchase.monthJune	0.0948	0.2000	0.4737	0.6357
life.firstpurchase.monthJuly	0.1044	0.4196	0.2487	0.8036
life.firstpurchase.monthAugust	0.1169	0.4481	0.2609	0.7942
life.firstpurchase.monthSeptember	0.1278	0.3905	0.3273	0.7434
life.firstpurchase.monthOctober	0.0919	0.4017	0.2287	0.8191
life.firstpurchase.monthNovember	0.1002	0.3887	0.2579	0.7965
life.firstpurchase.monthDecember	0.0975	0.4300	0.2268	0.8206
life.firstpurchase_lines	0.0163	0.0309	0.5276	0.5978
life.categorized_valueFourthQuartile	0.1243	0.2870	0.4330	0.6650
life.categorized_valueThirdQuartile	0.1434	0.2746	0.5223	0.6015

	Estimate	Std. Error	z-val	Pr(>|z|)
life.categorized_valueSecondQuartile	0.1125	0.3366	0.3342	0.7382
trans.delegation.BARCELONA	0.1034	0.1990	0.5195	0.6034
trans.delegation.BIZKAIA	0.1024	0.2913	0.3515	0.7252
trans.delegation.GUIPUZCOA	0.1195	0.3047	0.3923	0.6948
trans.delegation.MADRID	0.1027	0.2770	0.3707	0.7109
trans.delegation.PONTEVEDRA	0.1054	0.7323	0.1440	0.8855
trans.delegation.SEVILLA	0.1270	0.2558	0.4963	0.6197
trans.delegation.TARRAGONA	0.1119	1.0005	0.1118	0.9110
trans.delegation.VALENCIA	0.0723	0.3052	0.2370	0.8127
trans.delegation.ZARAGOZA	0.1448	0.5778	0.2506	0.8021
trans.same.zoneTRUE	0.1075	0.2009	0.5349	0.5927
trans.fp.transportTheir_means	0.0839	0.1263	0.6642	0.5065
trans.distance	-0.0055	0.0004	-14.1443	0.0000
trans.treatmentSilver	0.0997	0.2187	0.4560	0.6484
trans.marketOEM	0.1027	0.2060	0.4984	0.6182
trans.marketOTR	0.1136	0.0024	47.4970	0.0000
trans.marketSI	0.0991	0.2824	0.3511	0.7255
trans.storeYes	0.0592	0.1108	0.5337	0.5935
trans.digita_cliYes	0.0776	0.5329	0.1456	0.8843
trans.firstpurchase.monthFebruary	0.0957	0.2841	0.3367	0.7363
trans.firstpurchase.monthMarch	0.1494	0.3608	0.4140	0.6789
trans.firstpurchase.monthApril	0.0975	0.2939	0.3318	0.7401
trans.firstpurchase.monthMay	0.1040	0.2344	0.4437	0.6573
trans.firstpurchase.monthJune	0.1144	0.1962	0.5831	0.5599
trans.firstpurchase.monthJuly	0.1049	0.3468	0.3025	0.7623
trans.firstpurchase.monthAugust	0.1190	0.3894	0.3055	0.7600
trans.firstpurchase.monthSeptember	0.0985	0.3294	0.2990	0.7650
trans.firstpurchase.monthOctober	0.0573	0.3151	0.1817	0.8558
trans.firstpurchase.monthNovember	0.0958	0.3213	0.2980	0.7657
trans.firstpurchase.monthDecember	0.1085	0.3317	0.3270	0.7436
trans.firstpurchase_lines	0.0593	0.0308	1.9273	0.0539
trans.categorized_valueFourthQuartile	0.1055	0.1870	0.5641	0.5727
trans.categorized_valueThirdQuartile	0.1038	0.1940	0.5352	0.5925
trans.categorized_valueSecondQuartile	0.1166	0.1901	0.6133	0.5397

Table 7.2, Coefficients of the complete model with all covariates for the fifth cohort, where life indicates the variable is in the lifetime process and trans indicates the variable is in the transaction process

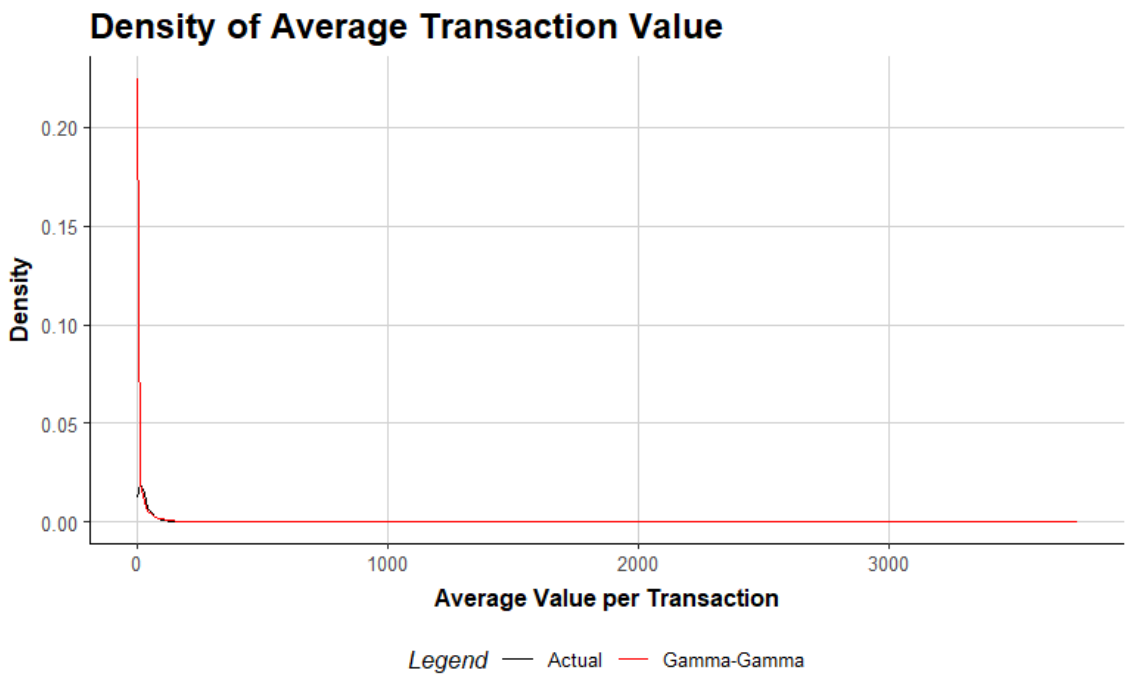


Figure 7.21, expected and observed density distributions of the average transaction value in the fifth cohort.

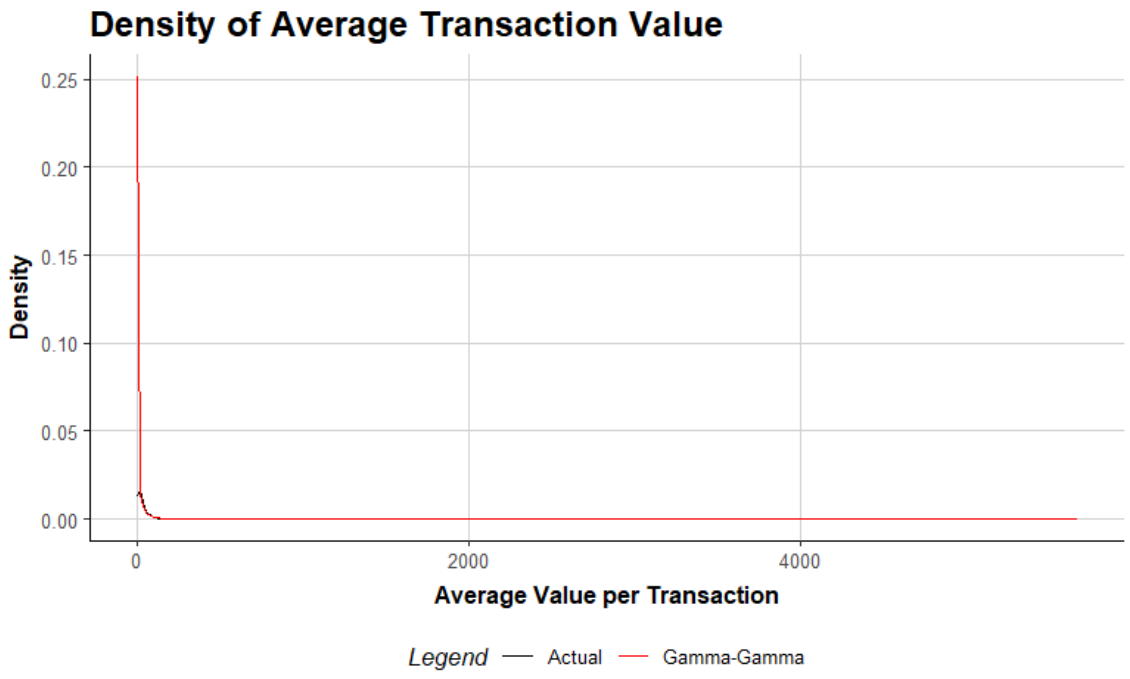


Figure 7.22, expected and observed density distributions of the average transaction value in the sixth cohort.