TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Fidelidad del asegurado usando técnicas de *machine learning*.

Autor: Patricia Fernandez Di Stefano

Director: Catalina Bolance Losilla

Departamento: Departamento de Econometría, Estadística y Economía Aplicada

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen

Se utilizaron métodos de *machine learning*, concretamente las NN-neural networks (redes neuronales) y el SVM support vector machine (máquina de soporte vectorial) para entrenar modelos con el objetivos de predecir la probabilidad de que un asegurado cancele su poliza de seguros de auto u hogar. También se estudia si la siniestralidad guarda relacion con esta cancelación.

Los resultados obtenidos en esta investigación ponen de manifiesto como las NN se pueden adaptar mejor a la problemática de la desproporción del estado de las pólizas, ya que las SVM tienden a favorecer a las pólizas vigentes por ser la clase mayoritaria. También se evidencia la importancia de la siniestralidad, las características contractuales y su histórico a la hora de modelizar la predicción de cancelar la póliza.

Palabras claves: Aprendizaje automatizado, ciencias actuariales, inteligencia artificial, modelos estadísticos, fidelización del cliente, redes neuronales artificiales, máquinas de soporte vectorial, retencion del asegurado.

Insured customer loyalty using machine learning techniques.

Abstract

Machine learning methods were used, specifically NN (neural networks) and SVM (support vector machine) to train models with the objective of predicting the probability that an client will cancel their auto or home insurance policy. It is also studied whether the claims rate is related to this cancellation.

The results of this research show how the NN can be better adapted to the problem of the disproportion of the state of the policies, since the SVM favors the not canceled policies because they are the majority class. The importance of repoting claims, the contractual characteristics and its history are also evidence when modeling the prediction of canceling the policy.

Keywords: *Machine learning, actuarial science, artificial intelligence, statistical models, customer loyalty, artificial neural networks (ANN), support vector machines (SVM), client retention.*

Clasificación AMS:

62XX Statistics

68TX Artificial intelligence

68WX Algorithms

INDICE

I. INTRODUCCION

La estadística abarca tantos campos diferentes de la vida como se pueda imaginar y plantear y resulta realmente interesante como es la base de distintos procesos como lo son los algoritmos de *machine learning* por su enfoque probabilístico, sin embargo, dicho algoritmo difiere ligeramente de la estadística en términos de énfasis y terminología.

El *machine learning (*aprendizaje automático), es un término que en la actualidad se escucha mucho y como su nombre indica, se puede decir que consiste en transformar una información a través de distintos algoritmos de *machine learning* a una nueva información, que según el interés que se desee, la maquina entrenara la informacion clasificándola de tal forma que podrá dar respuestas a una situación o problema a futuro. El *machine learning* puediera aportar resultados similares a los de algunos modelos estadísticos más tradicionales, aunque no los remplazaría.

Como se mencionó anteriormente, la estadística se puede aplicar a cualquier aspecto de la vida diaria, en este trabajo se estudia en el ámbito del seguro. En Europa es obligatorio contar con un seguro de vehículo que cubra por lo menos los daños a terceros. Con tantas opciones en el mercado actualmente la retención del cliente es de gran importancia para las compañías de seguro, por eso se desea estudiar la probabilidad que un asegurado cancele su póliza.

Por ser tan versátil las técnicas de *machine learning* supervisadas, son las estudiadas y utilizadas en este trabajo junto a sus retos algorítmicos. Concretamente, se empleará el modelo de *NN-Neural network* (redes neuronales) y *SVM-support vector machine* (máquinas de soporte vectorial) porque a través de sus algoritmos clasifica la información de entrada, según la de salida deseada, entrenando así a la máquina, por tanto para modelizar la retención del cliente se usa como salida deseada la información de polizas canceladas de una matriz de datos real.

El objetivo es modelizar, analizar y validar la retención del cliente de una compañía de seguros según las carateristicas contractuales de las pólizas de seguros de auto y de hogar. Así mismo, se hace énfasis en estudiar el efecto que tiene la siniestralidad, declarar o no siniestros, a la hora que un cliente cancele su póliza. Esta siniestralidad no se conoce a priori y tiende guardar relaciòn con el hecho de cancelar la póliza principalmente porque les podría aumentar la prima una vez se declaran siniestros por culpa del asegurado.

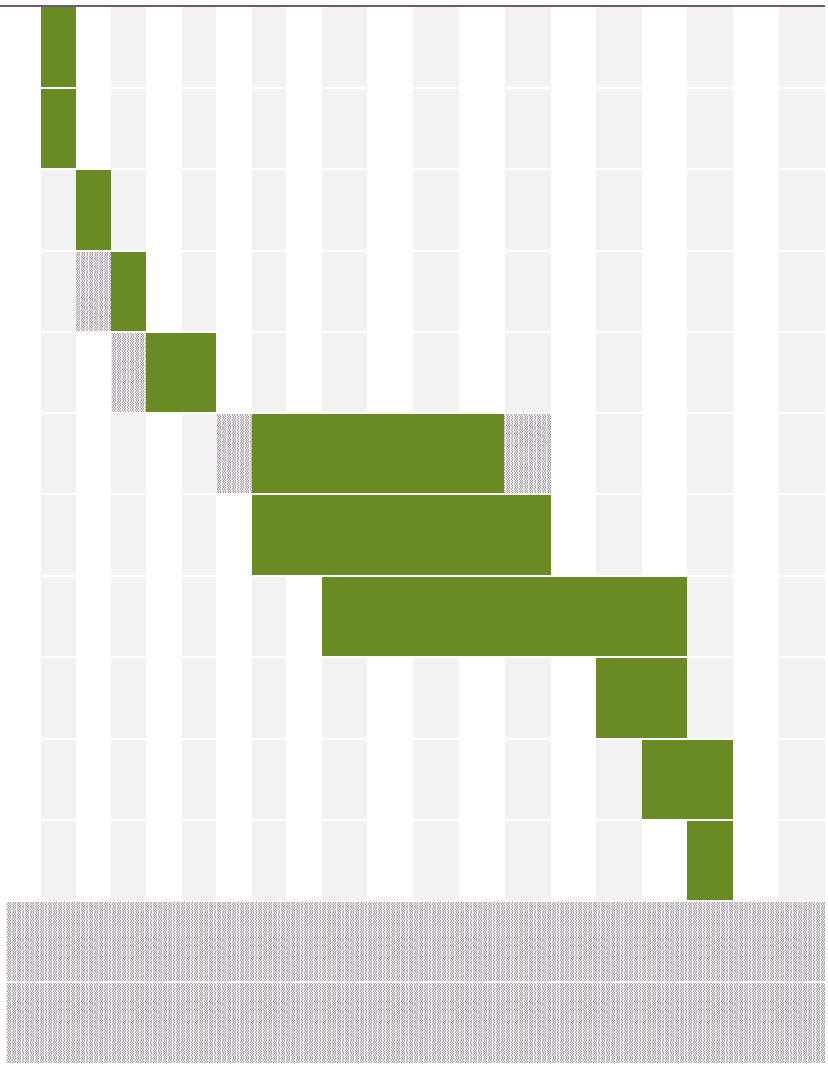
El trabajo consta de un apartado para explicar conceptos básicos del sector asegurador y profundizar en la renovación del cliente, posteriormente se describe la matriz de datos utilizada con un analisis univariante y bivariante de las variables, consecutivo está el apartado donde se

explican los metodos de *machine learning* utilizados y por consiguiente en el apartado posterior se evidencian los resultados de la aplicación de los metodos del estudio de la fidelización del cliente. Finalmente hay una pequeña discusión comparando los metodos y la conclusión.

El contenido del trabajo se organizó con los tiempos establecidos del siguiente cronograma:

Plan de acción

ACTIVIDAD	INICIO DEL PLAN	DURACIÓN DEL PLAN
Portada y AMS	2	1
Resumen y palabras claves	2	1
Índice y estructura	3	1
Introducción	3	1
Depuración base de datos	4	2
Investigación	7	8
Cuerpo del trabajo	8	7
Aplicación de métodos	10	8
Resultados	16	2
Discusión	17	2
Conclusiones	18	1
Bibliografía	1	20
Anexos: códigos	1	20



Por último, el recurso informático que se utiliza es R-studio por ser uno de los programas más trabajados a lo largo del grado y poseer un amplio repertorio de paquetes estadísticos, de *machine learning* y de visualización gráfica.

II. SECTOR ASEGURADOR

Al ser el objetivo estudiar la probabilidad de que un asegurado cancele su póliza es fundamental conocer ciertos conceptos básicos del sector asegurador, así como la importancia de esta probabilidad.

2.1 CONCEPTOS BÁSICOS

- **Ciencias actuariales:** Disciplina en la que se evalúan los riesgos que soportan las compañías de seguros mediante la aplicación de técnicas estadísticas y matemáticas.

- **Contrato de seguro**: (Ley 50/1980) es un contrato que obliga a la compañía aseguradora mediante el cobro de una prima y para el caso de que se produzca un siniestro, cuyo riesgo está cubierto a indemnizar dentro de los limites pactados, el daño producido al asegurado (cliente).

- **Riesgo**: Es el daño potencial que puede surgir por un proceso presente o suceso futuro, es la posibilidad de que un peligro pueda llegar a materializarse.

- **Póliza**: Documento contractual que la compañía aseguradora entrega al cliente o tomador del seguro, mediante el cual se declara formalmente la existencia del contrato de seguro y las condiciones en que se ha pactado.

- **Prima**: Precio del seguro. Es la cantidad de dinero que el cliente tomador paga para que la compañía aseguradora indemnice en caso de siniestro.

- **Prima pura:** Valor esperado del coste real del riesgo que la compañía aseguradora asume.

- **Siniestro**: Evento perjudicial cuyo riesgo está contemplado en la póliza y que se ha producido durante la cobertura de esta.

- **Tipos de seguro:** Existen distintos tipos de clasificaciones, entre ellas el objeto asegurado:
    - Seguros personales: El objeto asegurado es una persona, el pago de la prestación depende de su existencia, salud o integridad. Entre ellos están el seguro de salud, accidentes y el seguro de vida
    - Seguros de daños: Seguros cuyo fin principal es reparar el daño sufrido como consecuencia del siniestro en el patrimonio del tomador del seguro.

Entre ellos están: seguros de no vida, seguro del automóvil, del hogar, etc.

2.2 FIDELIDAD DEL CLIENTE

La fidelización o retención de clientes y la predicción de abandono han ganado importancia en múltiples de sectores financieros debido a la alta rotación de clientes y competitividad.

La fidelización del cliente es una cuestión importante de estudiar por sus múltiples efectos que puede tener en una compañía de seguros, estos efectos pueden ser, tanto económicos, como de marketing, ya que, como dicen algunos expertos “la fidelización o retención es como la moneda de la compañía” porque se pudiera conocer el beneficio aproximado por la cantidad de clientes activos.

Perder un cliente pudiera significar perder más clientes y si son con ratios de siniestralidad baja, probablemente conducirá a una reducción de beneficios a la compañía.

El estudio de la fidelidad puede dar a conocer los perfiles de cliente más rentable y la predicción de esta retención pronostica, que cantidad de clientes son más probables a trasladarse a otra empresa competitiva. Cancelar la póliza suele estar relacionado al hecho de declarar siniestros, porque la prima puede aumentar de un periodo al siguiente por el hecho de poseer siniestros declarados por culpa del asegurado.

También es fundamental conocer el estado actual de la cartera de clientes, ya que, gracias a esa información se pueden implementar distintos planes de acción, como mejorar la calidad de servicio al cliente o beneficios de estar en la compañía.

Tradicionalmente en las ciencias actuariales se utilizan modelos estadísticos clásicos como el lineal (ML) o lineal generalizado (GLM), pero, existen otras metodologías como las de *machine learning,* que resultan interesantes de implementar y son las utilizadas en este trabajo.

Es indispensable realizar un estudio y minería de datos para una predicción lo más precisa posible, hay que tener en cuenta que la fidelidad del cliente se estudia a través del estado de las pólizas, están pueden estar vigentes (se ha retenido al cliente) o canceladas (se perdió al cliente).

Otro aspecto que considerar es que las pólizas canceladas representan un porcentaje mucho menor que las vigentes, por consiguiente, se suele decir que las clases no están balanceadas. Debido a este desbalance entre clases puede existir una mala predicción de *machine learning* favoreciendo a la clase mayoritaria, para evitar esto, se trata de que las categorías de las variables explicativas sean lo más balanceadas posibles

.

III. DESCRIPCIÓN DE LOS DATOS

3.1 ORIGEN

La base de datos utilizada para determinar la probabilidad de que un asegurado cancele su fue proporcionada por la tutora, Catalina Bolance. Se trata de una base de datos no normalizada y como clave tenía el número de póliza y el identificador del cliente. Cuenta con pólizas de autos y de pólizas de hogar.

3.2 ESTRUCTURA BÁSICA LA MATRIZ DE DATOS:

La estructura de la matriz de datos se compone de 33 variables (columnas) y de 745569 observaciones (filas) de las cuales 538757 son pólizas de auto y 206812 pólizas de hogar.

A continuación, en las siguientes tablas, se muestran las variables originales con su tipología, descripción y al grupo de pólizas al que pertenece cada variable.

- **Estructura de las variables en común entre los dos grupos de pólizas**

Variable	Tipo	Descripción
nunpol_life_risk	Numérica	Número de pólizas de vida
nunpol_life_saving	Numérica	Número de pólizas de ahorros
nunpol_accidents	Numérica	Número de pólizas contra accidentes
nunpol_InsRetPlan	Numérica	Número de pólizas de plan de retiro
nunpol_other	Numérica	Número de otras pólizas
sex_customer	Categórica	Sexo del cliente
Age_client	Numérica	Edad del cliente
Client_Seniority	Numérica	Antigüedad del cliente
Policy_PaymentMethod	Categórica	Método de pago de la póliza
previous_to_last_premium_paid	Numérica	Pago por la penúltima prima
last_premium_paid	Numérica	Pago por la ultima prima
dif_current_previous	Numérica	Diferencia entre la prima actual y la previa
dif_current_first	Numérica	Diferencia entre la prima actual y la primera
policy_group	Numérica	Grupo de auto u hogar
policy_status_at_t	Categórica	Estado de la póliza

Tabla 3.1. Estructura de los datos auto y hogar

- **Estructura pólizas de auto**

Variable	Tipo	Descripción
age_of_car_M	Numérica	Antigüedad del vehículo

Car_Years1stDriverLicense_M	Numérica	años del permiso de conducir del primer conductor
Car_number_of_seats_M	Numérica	Número de asientos del vehículo
Car_power_M	Numérica	Potencia del vehículo
car_bonus_M	Categórica	Bono
car_type_M	Categórica	Tipo de vehículo
contracted_guarantee_M	Categórica	Garantía contratada
nclaims_md_ins_C	Numérica	Número de siniestros daño material por culpa del asegurado
nclaims_bi_ins_C	Numérica	Numero siniestros de lesión corporal por culpa del contrario
nclaims_md_con_C	Numérica	Número de siniestros daño material por culpa contrario
nclaims_bi_con_C	Numérica	Numero siniestros de lesión corporal por culpa del contrario
age_of_car2_M	Numérica	Antigüedad del vehículo 2
Policy_numSupplements	Numérica	Numero de suplementos de la póliza

Tabla 3.2. Estructura de los datos auto

En resumen, es una matriz de 28 variables y 538757 observaciones.

- **Estructura póliza de hogar**

Variable	Tipo	Descripción
Insuredcapital_content_H	Numérica	Contenido de capital asegurado
Insuredcapital_continent_H	Numérica	Continente capital asegurado
HomeType_H	Categórica	Tipo de hogar
nclaims_home_ins_C	Numérica	Número de siniestros hogar culpa del asegurado
nclaims_home_con_C	Numérica	Número de siniestros hogar culpa del contrario

Tabla 3.3. Estructura de los datos hogar

La matriz de las pólizas de hogar está compuesta por 20 variables y 206812 observaciones.

3.3 PREPROCESAMIENTO

El preprocesamiento de los datos o *prepeocesing* en inglés, es una técnica de la minería de datos. Consiste en la limpieza y tratamiento de valores atípicos o inexistentes de las variables, evita generar errores y asegura que los datos sean completos y precisos para futuras acciones. Es importante transforma los datos sin procesar en un formato que sea comprensible, ya que, los datos del mundo real pueden ser incompletos (poseen missings), inconsistentes (posibles *outliers*) o simplemente difíciles de interpretar. Este proceso prácticamente se dividió para tratar los 4 problemas siguientes:

3.3.1 Depuración

Se agruparon variables que aportaran la misma información en una sola con el fin de simplificar la base de datos, estas nuevas variables son:

- El número de siniestros [*totalclaims*] es la suma de los números de siniestros de todos los tipos de siniestros que tuvo la póliza por culpa del asegurado. Se eliminan los siniestros por culpa del contrario ya que estos no afectan directamente el precio de las primas.

  [*totalclaims*] = nclaims_bi_ins_C+ nclaims_md_ins_C + nclaims_home_ins_C

- El número total de pólizas [*totalpol*] se agrupan en una categoría que luego se transforma en binaria el número de todos los otros tipos de póliza adicional a la del registro.

  [*totalpol*] = nunpol_accidents+ nunpol_InsRetPlan+ nunpol_life_risk+ nunpol_life_saving+ nunpol_other

3.3.2 Recodificación

Se recodifican las variables según su valor original a un código binario, es decir a un sistema numérico de dos dígitos 0 y 1, esto simplificara el uso de los métodos de predicción. Para crear clases lo más balanceadas posibles algunas variables con más de 2 clases se simplifican en una variable binaria.

Variables respuestas para ambas pólizas:

- El estado de la póliza [policy_status_at_t]: Vigente = 1 y Anulada = 0

Variables explicativas para ambas pólizas:

- Sexo del cliente [sex_customer]: Female=1 y Male=0
- Grupo de la poliza [policy_group]: hogar= 1 y auto =0
- Método de pago [Policy_PaymentMethod]: método anual “A” = 1 y resto de métodos "S"="T"="U"= 0
- El número total de pólizas [*totalpol*] de forma que 0 es que no tiene ninguna póliza adicional a la del registro y 1 es que tiene más de 1 póliza
- Probabilidad de declarar siniestros [pclaims]: más de 1 siniestro declarado =1 y ningún siniestro declarado = 0

- Diferencia entre la prima actual y la previa [dif_current_previous]: 0 si la diferencia es positiva, 1 si la diferencia es negativa.
- Diferencia entre la prima actual y la primera [dif_current_first]: 0 si la diferencia es positiva, 1 si la diferencia es negativa.
- El número de suplementos asociados a la póliza [Policy_numSupplements]: toma valor 1 si tiene suplementos y 0 si la póliza no posee suplementos.

Variables explicativas para póliza de auto:

- Segundo conductor [Car_2ndDriver_M]: YES=1 y NO=0
- Garantía contratada [contracted_guarantee_M]: THIRD PARTY (a terceros) = 1, ALL RISKS WITHOUT FRA (todo riesgo sin franquicia) = ALL RISKS WITH FRA (todo riesgo con franquicia) =0
- Tipo de auto [car_type_M]: "TOURISM"=1,"MINIVAN"= ”ALL TERRAIN"=0
- Tipo de combustible [Fuel_Type_M]: “Gasoline"=1,"Diesel"= "Hybrid"=0
- Bonus [car_bonus_M]: 0 indica que no tiene bono y el 1 indica que tiene bono del 20% al 55% de bono en la póliza.
- Número de asientos [Car_number_of_seats_M]: “5” = 1, “1” =”2” =”3” =”4” =”6” =”7” =”8” =”9” = 0

Variables explicativas para póliza de hogar:

- Tipo de hogar [HomeType_H]: "PI"=1, "AT"="PB"="RU"="UA"="UF"=0

Por último, se categorizan las variables numéricas según por clúster, agrupando las categorías de una variable por su similitud.

Variables explicativas para póliza de auto:

- Valor del auto [value_of_car_M]: Se agrupan los 44 intervalos en 3 categorías.
  - [value_of_car_9k]: Es una variable binaria que toman valor 1 los autos de €0 hasta los €9000.
  - [value_of_car_M]: es otra variable binaria que toman valor 1 los autos de €9000 hasta €23000, significando que la variable de referencia (0 en las dos anteriores) son los autos superiores a los €23000.

- Potencia del vehículo [Car_power_M]: Se transforma en un variable binaria de forma que las pólizas con autos con potencias menores a los 128CV (caballos de vapor) toman valor 1 y el resto 0.

3.3.3 Missings

Al tratarse de una base de datos de gran tamaño se eliminaron todos aquellos registros que no tienen información en alguna de sus variables, para evitar interferencia en las predicciones. Eliminándose en total solo el 0.05% de las observaciones originales.

3.3.4 Outliers

En la variable [*previous_to_last_premium_paid*] para las pólizas de auto, existen 5 pólizas con primas superiores a los €150000 cuyos valores del vehículo era inferiores a los €3000, lo cual no parecía ser correcto así que, se eliminaron los registros.

También se eliminan los registros cuyas diferencias absolutas entre la prima actual y la anterior [dif_current_previous] y las diferencias absolutas entre la póliza actual y la primera [dif_current_first] superase los €1000.

Estos valores atípicos representaban aproximadamente un 0.1% de los datos. Finalmente queda un total de **463030 observaciones y 23 variables paran las pólizas de auto** y **172708 observaciones y 13 variables para las de hogar.**

3.4 ANÁLISIS DESCRIPTIVO PÓLIZAS DE AUTO

Como se mencionó en el apartado II, conocer las variables que se desean estudiar, sus tendencias y frecuencias es fundamental para poder ejecutar los métodos predictores y que estos sean lo más óptimos posible, por ese motivo se realiza un análisis univariante y bivariante.

3.4.1 ANÁLISIS DESCRIPTIVO UNIVARIANTE

A continuación, se presenta el análisis univariante para las 6 variables cuantitativas y las 15 variables categóricas de las pólizas de auto.

3.4.1.1 Variables cuantitativas

- **Variable Age_client**

Esta variable numérica informa la edad del cliente poseedor de la póliza en años. Es una variable que toma valores enteros entre 18 y los 99 años.

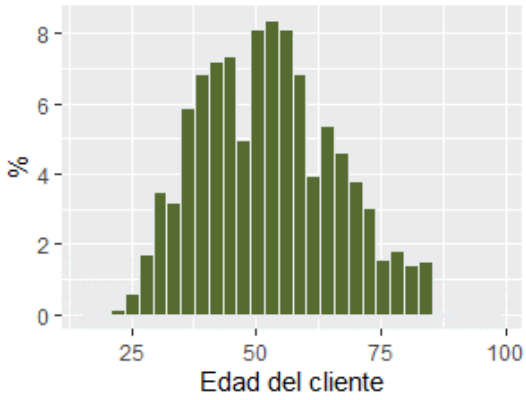


Figura 3.1 Histograma edad

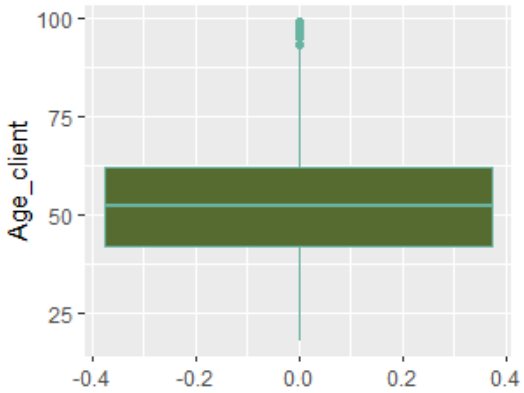


Figura 3.2 Boxplot edad

Min.	1stQu.	Median	Mean	3rdQu.	Max.
18.00	42.00	52.00	52.28	62.00	99.00

Tabla 3.4 Estadísticos edad

En las gráficas se aprecia que el 50% de los clientes se encuentran entre los 42 y 62 años.

- **Variable Client_Seniority**

Esta variable permite conocer la antigüedad en años del cliente poseedor de la póliza en la compañía. A continuación, se muestra una tabla de frecuencias y porcentajes para poder darnos una idea de esta antigüedad.

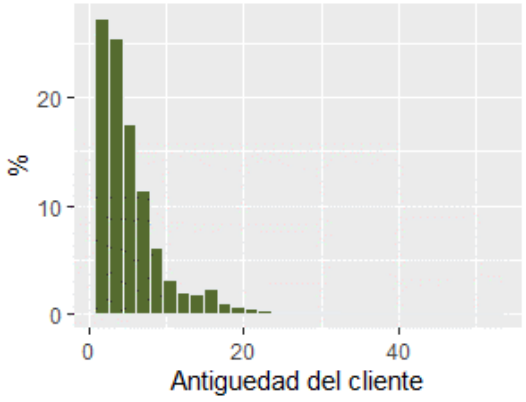


Figura 3.3 Histograma de la antigüedad

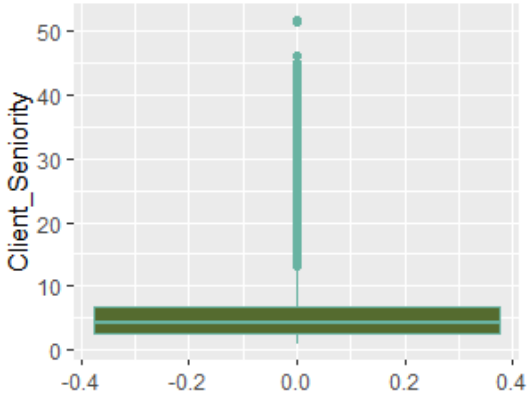


Figura 3.4 Boxplot de la antigüedad

Min.	1stQu.	Median	Mean	3rdQu.	Max.
1.002	2.491	4.115	5.425	6.793	51.759

Tabla 3.5 Estadíscos de la antigüedad

Varía desde 1 año hasta 51 años de antigüedad, aunque este valor parece ser atípico, ya que la media son 5 años y el tercer cuartil son casi 7 años.

- **Variable previous_to_last_premium_paid**

El pago previo a la última prima pagada lo que es igual a la penúltima prima pagada es una variable continua que toma valores desde €3 hasta €65445.

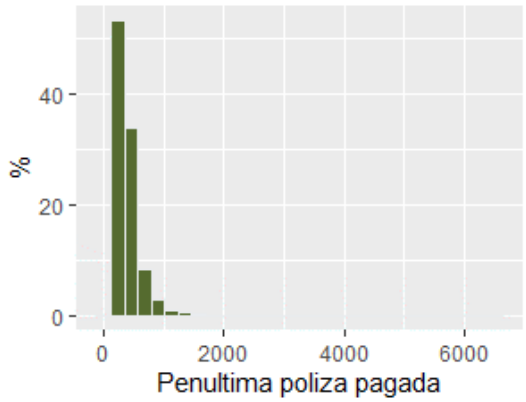


Figura 3.5 Histograma de penúltima prima pagada

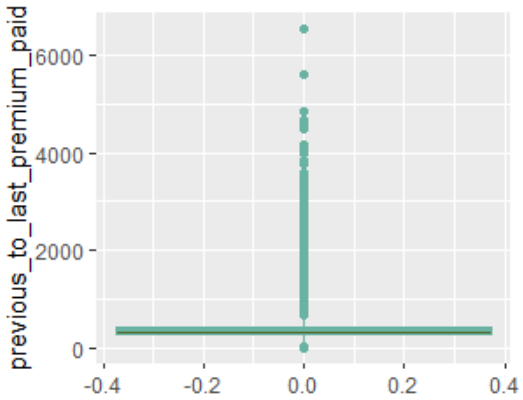


Figura 3.6 Boxplot de penúltima prima pagada

Min.	1stQu.	Median	Mean	3rdQu.	Max.
2.8	267.1	329.1	385.6	434.9	6544.8

Tabla 3.6 Estadíscos de penúltima prima pagada

El monto medio pagado por la penúltima prima es de €385 y el 75% de los datos se concentran en montos menores de €435.

- **Variable last_premium_paid**

Se refiere a cantidad pagada por la ultima prima, es una variable continua que abarca valores desde los €4 hasta los €6445.

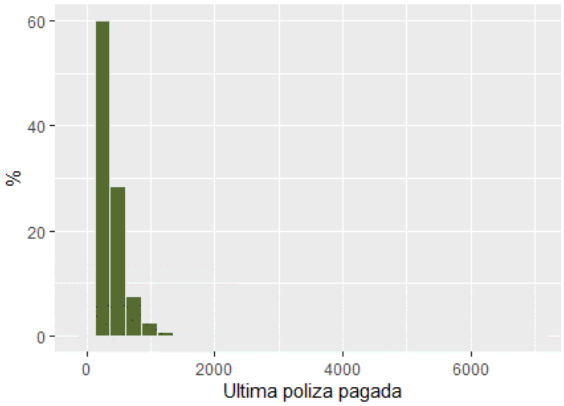


Figura 3.7 Histograma de última prima pagada

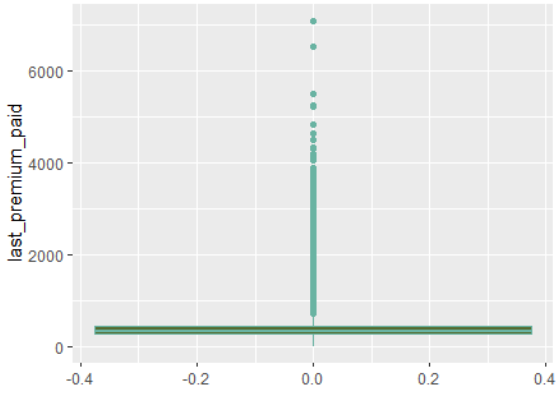


Figura 3.8 Boxplot de última prima pagada

Min.	1stQu.	Median	Mean	3rdQu.	Max.
4.04	269.83	333.07	391.82	443.89	6544.80

Tabla 3.7 Estadísdicos de ultima prima pagada

Sigue una distribución similar a la variable anterior, el 75% de los datos se concentran entre los €4 hasta los €444.

- .**Variable age_of_car_M**

La antigüedad del vehículo es un variable discreta que va desde los 0 hasta los 88 años.

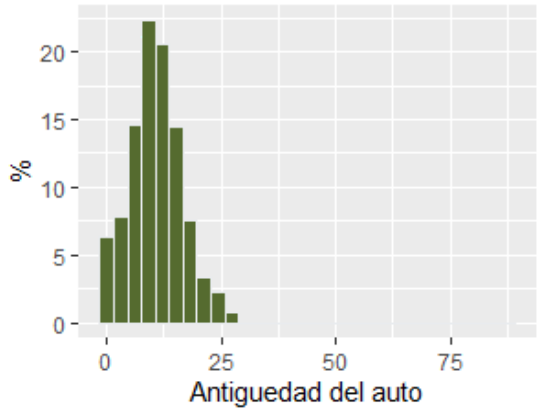


Figura 3.9 Histograma antigüedad del auto

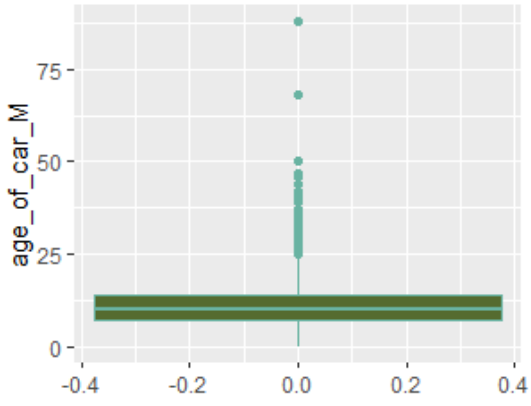


Figura 3.10 Boxplot antigüedad del auto

Min.	1stQu.	Median	Mean	3rdQu.	Max.
0.00	7.00	10.00	10.57	14.00	88.00

Tabla 3.8 Estadísticos antigüedad del auto

En promedio los autos de las pólizas tienen 10 años y existen excepciones que superan los 14 años de antigüedad.

- **Variable Car_Years1stDriverLicense_M**

Los clientes poseedores de las pólizas pueden poseer desde 0 hasta 79 años con sus permisos de conducir.

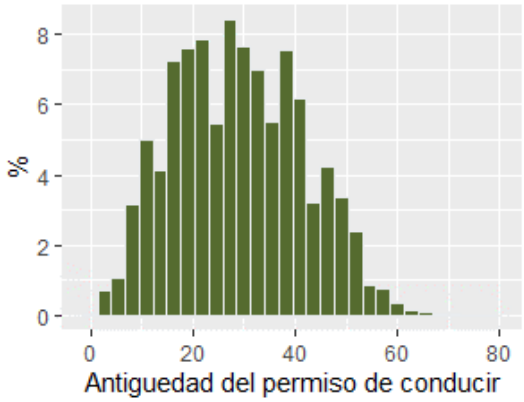


Figura 3.11 Histograma antigüedad del permiso de conducir

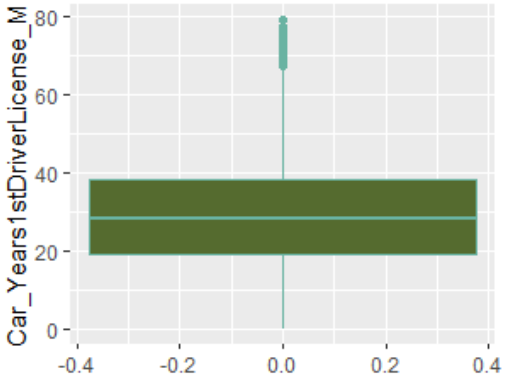


Figura 3.12 Boxplot antigüedad del permiso de conducir

Min.	1stQu.	Median	Mean	3rdQu.	Max.
0.00	19.00	28.00	28.87	38.00	79.00

Tabla 3.9 Estadísticos de antigüedad del permiso de conducir

En promedio tienen 28 años con sus permisos de conducir y los datos están concentrados entre los 20 y 40 años y existe una dispersión hasta los 78 años de antigüedad del permiso.

3.4.1.2 Variables categóricas

- **Variable sex_customer**

Esta variable informa del sexo del poseedor de la póliza, es binaria de forma que el 1 representa al sexo femenino y el 0 al masculino.

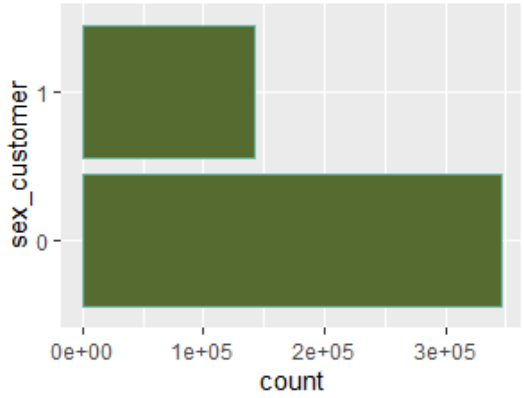


Figura 3.13 Diagrama de barras del sexo

sex_customer

	Frequency	Percent
0	326524	70.52
1	136506	29.48
Total	463030	100.00

Tabla 3.10 Frecuencias del sexo

Se puede apreciar en la tabla y gráficamente que el 70% son de sexo masculino y el 30% restante de sexo femenino.

- **Variable Policy_PaymentMethod**

Esta variable se transforma en binaria con el fin de crear clases más equilibradas, si el método de pago es 1 el pago es anual y si el 0 al resto de tipologías de pago, S, T y U.

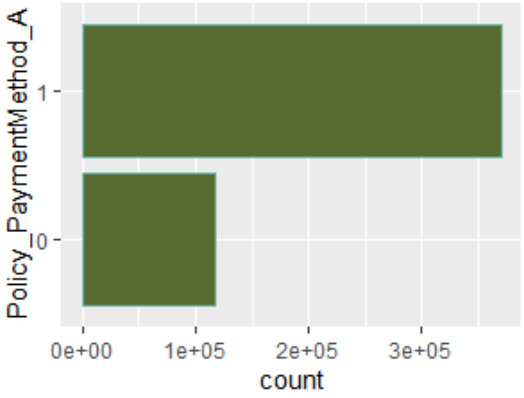


Figura 3.14 Diagrama de barras del método de pago

Policy_PaymentMethod_A

	Frequency	Percent
0	112187	24.23
1	350843	75.77
Total	463030	100.00

Tabla 3.11 Frecuencias del método de pago

El 75% de las pólizas son pagadas de forma anual, el 25% restante de forma S, T y U.

- **Variable policy_status_at_t**

El estado de la póliza es la variable de estudio posteriormente, indica el estado en el que se encuentra la póliza, es una variable binaria, el 1 representa a las pólizas vigentes y el 0 a las anuladas.

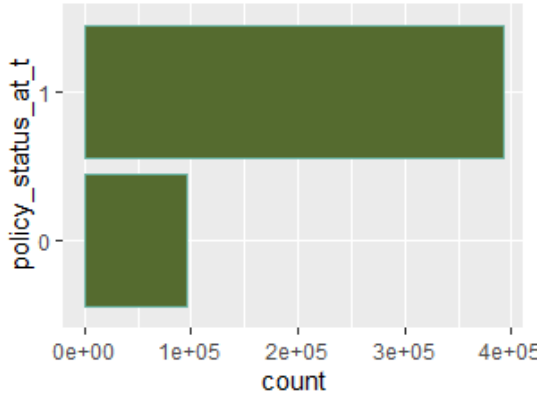


Figura 3.15 Diagrama de barras del estado de la póliza

policy_status_at_t

	Frequency	Percent
0	95729	19.63
1	391881	80.37
Total	487610	100.00

Tabla 3.12 Frecuencias del estado de la póliza

Se evidencia la desproporción entre las dos clases, las pólizas vigentes representan un 80% de todas las pólizas de auto y el 20% las canceladas.

- **Variable pclaims**

La probabilidad de declarar siniestro es la siniestralidad, el 0 representa 0 siniestros declarados y el 1 representa que ha sufrido algún siniestro.

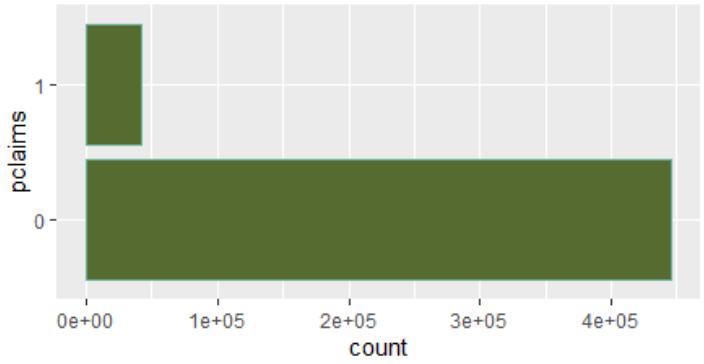


Figura 3.16 Diagrama de barras de pclaims

pclaims

	Frequency	Percent
0	446069	91.481
1	41541	8.519
Total	487610	100.000

Tabla 3.13 Frecuencias de la probabilidad de siniestro

La probabilidad de declarar siniestro en las pólizas del auto es de 8.5%.

- **Variable contracted_guarantee_M**

Esta variable indica el tipo de garantía contratada, se transformó a binaria por poseer categorías poco representadas, el 1 se refiere a pólizas con garantías contra terceros y el 1 a pólizas a todo riesgo con y sin franquicia.

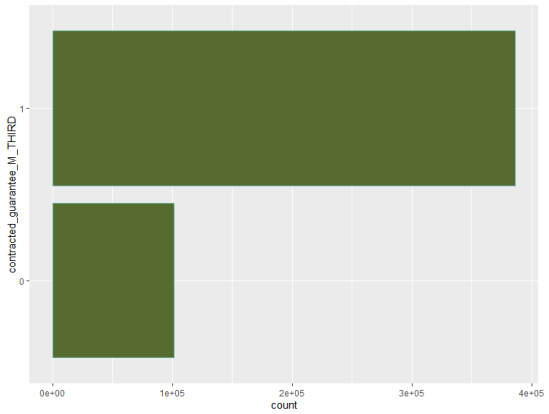


Figura 3.17 Diagrama de barras de garantía contratada

contracted_guarantee_M_THIRD

	Frequency	Percent
0	101194	20.75
1	386416	79.25
Total	487610	100.00

Tabla 3.14 Frecuencias de garantía contratada

Las pólizas que poseen una garantía contratada contra terceros representan un 79% y las pólizas con garantía a todo riesgo representan un 30%.

- **Variable totalpol**

La variable total de pólizas es una variable binaria e indica si el cliente poseedor de la póliza tiene asociada otras pólizas o no, el 0 representa que no es poseedor de ninguna otra póliza y el 1 que al menos posee una póliza más.

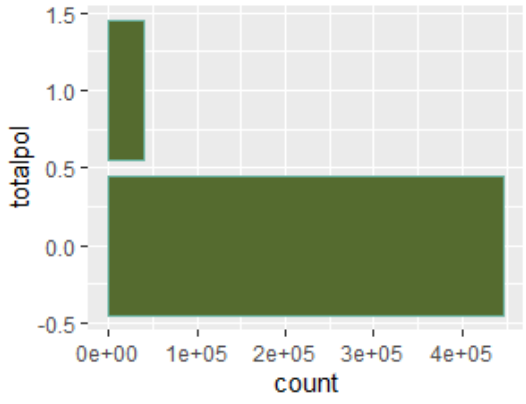


Figura 3.18 Diagrama de barras de pólizas adicionales

totalpol

	Frequency	Percent
0	447108	91.694
1	40502	8.306
Total	487610	100.000

Tabla 3.15 Frecuencias de pólizas adicionales

El 92% de las pólizas no posee otra póliza asociada y el 8% si tiene asociadas alguna otra póliza.

- **Variable dif_current_previous**

Esta variable indica la diferencia pagada entre la prima actual y la previa, es una variable que toma valor 0 si la diferencia es negativa (la previa es superior a la actual) y valor 1 si la diferencia es positiva (la prima actual es superior a la previa).

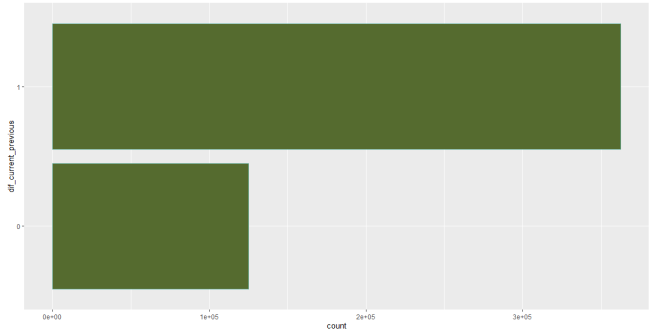


Figura 3.19 Diagrama de barras diferencia prima actual y previa

dif_current_previous

	Frequency	Percent
0	125133	25.66
1	362477	74.34
Total	487610	100.000

Tabla 3.16 Frecuencias diferencia prima actual y previa

El 26% de las primas actuales disminuyeron con respecto a la anterior y el 74% aumento.

- **Variable dif_current_first**

La diferencia entre la primera prima y la última prima pagada es una variable que toma valor 0 si la diferencia es negativa (la primera es superior a la última) y valor 1 si la diferencia es positiva (la última prima actual es superior a la previa).

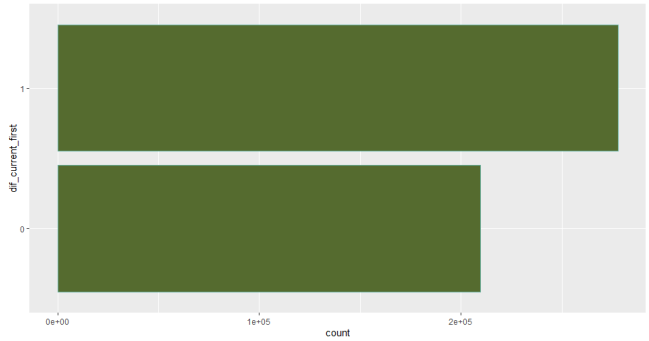


Figura 3.20 Diagrama de barras diferencia prima actual y primera

dif_current_first

	Frequency	Percent
0	209785	43.02
1	277825	56.98
Total	487610	100.000

Tabla 3.17 Frecuencias diferencia primas actual y primera

El 43% de las ultimas primas pagadas disminuyeron con respecto a la primera y el 57% aumento.

- **Variable Policy_numSupplements**

Los de suplementos asociados a la póliza es una variable binaria que toma valor 0 si la póliza no tiene ningún suplemento y valor 1 si tiene más de un suplemento.

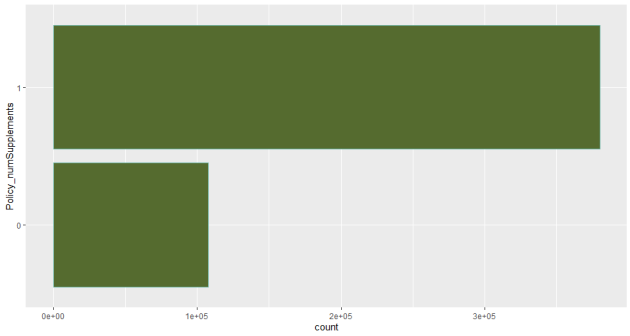


Figura 3.21 Diagrama de barras de suplementos

Policy_numSupplements

	Frequency	Percent
0	107692	22.09
1	379918	77.91
Total	487610	100.000

Tabla 3.18 Frecuencias de suplementos

Se observa que las pólizas que poseen suplementos representan un 78% y las que no tienen un 22%.

- **Variable car_type_M**

Esta variable binaria indica el tipo de uso al que está destinado el vehículo, si el auto es de tipo turismo toma valor 1 y si no, es de tipo miniván o todo terreno y toma valor 0.

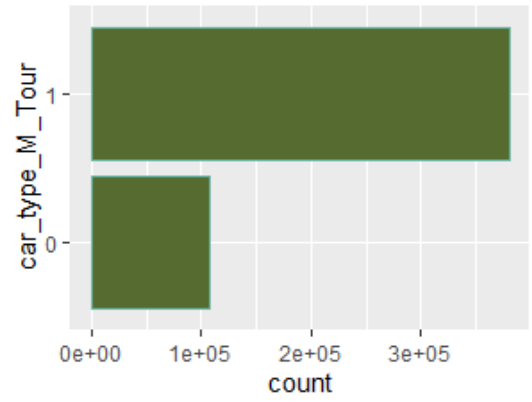


Figura 3.22 Diagramas de barras de tipo de auto turismo

car_type_M_Tour

	Frequency	Percent
0	107937	22.14
1	379673	77.86
Total	487610	100.00

Tabla 3.19 Frecuencias de tipo de coche Tour

El 78% de las pólizas poseen un auto de tipo turismo, el cual está destinado al uso particular, el 22% restante son miniván destinadas al transporte de mercancía ligera (hasta 3500kg) y de todo terreno, destinados a un uso intensivo en condiciones adversas.

- **Variable Car_2ndDriver_M**

La variable informa si en la póliza consta la existencia o no de un segundo conductor para el auto asegurado. Tomando valor 1 si posee segundo conductor y 0 si no posee segundo conducto.

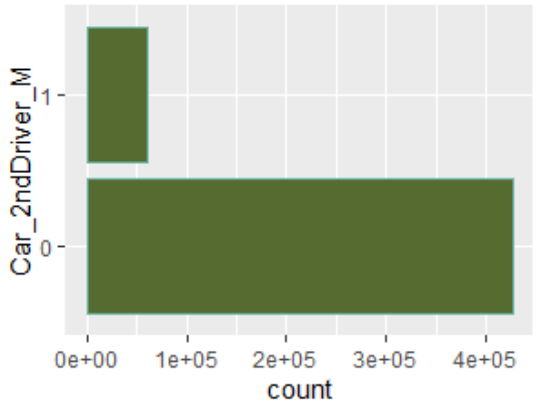


Figura 3.23 Diagrama de barras de 2do conductor

Car_2ndDriver_M

	Frequency	Percent
0	426589	87.49
1	61021	12.51
Total	487610	100.00

Tabla 3.20 Frecuencias de 2do conductor

Se observa que el 87.5% de las pólizas no tienen segundo conductor por tanto el 12.5% si tienen segundo conductor informado.

- **Variable Fuel_Type_M**

Los vehículos de las pólizas se pueden clasificar según el tipo de combustible que usan, esta variable indica si es de Diesel tomando el valor 1 y de gasolina o híbrido tomando el valor 0.

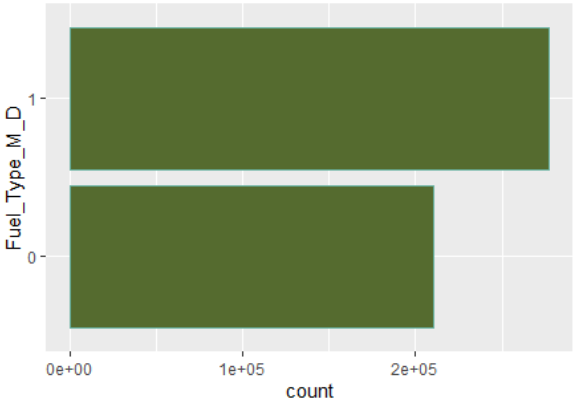


Figura 3.24 Diagrama de barras de tipo de combustible

Fuel_Type_M_D

	Frequency	Percent
0	210243	43.12
1	277367	56.88
Total	487610	100.00

Tabla 3.21 Frecuencias de tipo de combustible

Resulta que un 57% de los autos de las pólizas utilizan combustible Diesel y el resto mayoritariamente de gasolina junto a un 0.1% que es hibrido.

- **Variable Car_number_of_seats_M**

El número de asientos es una variable que se transformó in binaria indica si el auto tiene 5 asientos o no, toma el valor 1 para las pólizas con autos de 5 asientos y 0 para los autos con más o menos de 5 asientos.

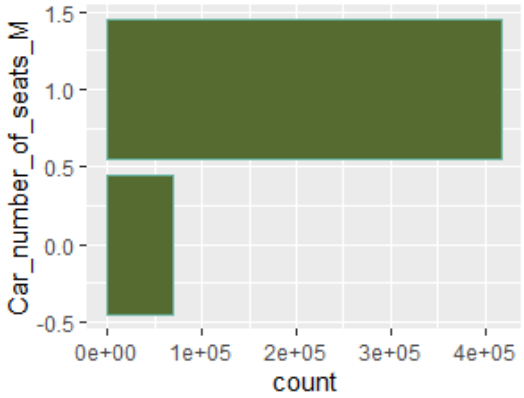


Figura 2.25 diagrama de barras de Nº de asientos del vehículo

Car_number_of_seats_M

	Frequency	Percent
0	70245	14.41
1	417365	85.59
Total	487610	100.00

Tabla 3.22 Frecuencias del Nº de asientos del vehículo

Se aprecia que el 85% de los autos de las pólizas poseen 5 asientos y el 15% restante pose 1,2,3,4,6,7,8 o 9 asientos.

- **Variable Car_bonus**

Por último, las pólizas de auto pueden verse beneficiadas por un bono o no, las pólizas que poseen un bono toman el valor 1 y las que no 0.

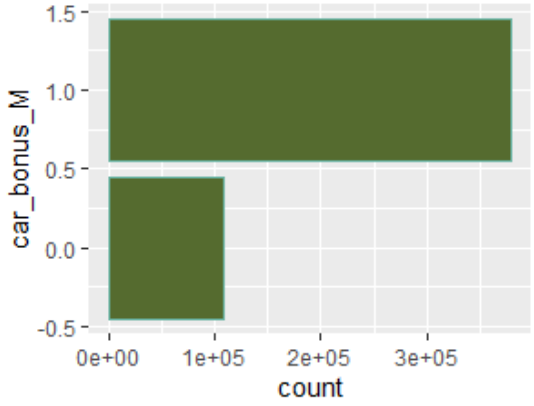


Figura 3.26 Diagrama de barras de bonos del vehículo

car_bonus_M

	Frequency	Percent
0	107692	22.09
1	379918	77.91
Total	487610	100.00

Tabla 3.23 Frecuencias de bonos del vehículo

El 22% de las pólizas no se benefician de ningún bono y el 78% restante se beneficia de bonos que van desde el 20% hasta el 55%.

- **Variable value_of_car_M**

Las pólizas cuyos autos tienen un valor entre €9000 y €22000 toman valor 1, y si son menores a 9 mil euros o superiores a 22 toman valor 0.

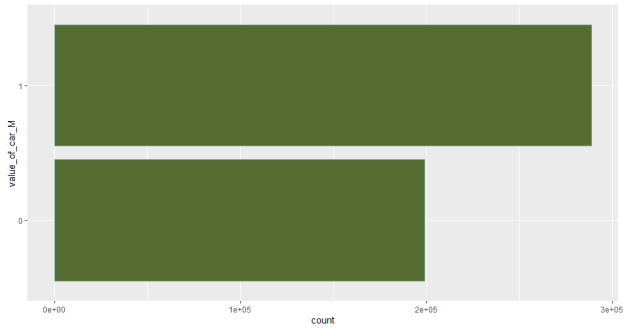


Figura 3.27 Diagrama de barras valor del auto 9-22k

value_of_car_M

	Frequency	Percent
0	198939	40.80
1	288671	59.20
Total	487610	100.000

Tabla 3.24 Frecuencias del valor del auto 9-22k

EL 59% de las pólizas aseguran autos de €9000 hasta €22000.

- **Variable value_of_car_9k**

Esta variable binaria indica si los autos asociados a la póliza tienen un valor inferior a los €9000, de lo contario toman valor 0. Si una observación toma valor 0 en esta variable y en la anterior significa que el auto tiene un valor superior a los €22000.

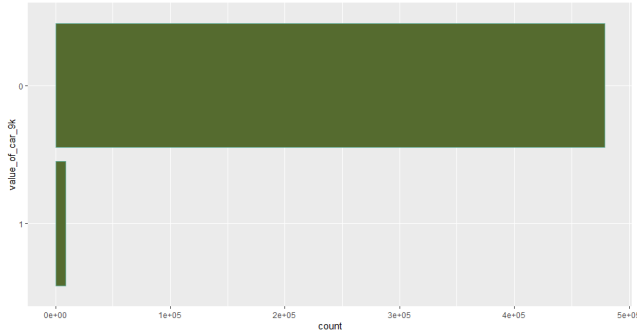


Figura 3.28 Diagrama del valor del auto hasta 9k

value_of_car_9k

	Frequency	Percent
0	8873	1.82
1	478737	98.18
Total	487610	100.000

Tabla 3.25 Frecuencias del valor del auto hasta 9k

Se observa que solo el 1,8% de las pólizas poseen autos de 0 a €9000 y que el 39% que tiene 0 en ambas variables son los autos de más de €22000.

- **Variable Car_power_M**

Esta variable indica la potencia del vehículo, si es inferior a los 128CV toma valor 1 y si es superior valor 0.

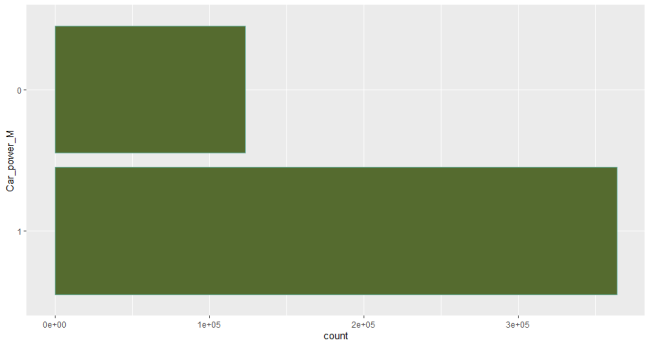


Figura 3.29 Diagrama de barras de la potencia del auto

Car_power_M

	Frequency	Percent
0	123499	25.33
1	364111	74.67
Total	487610	100.000

Tabla 3.26 Frecuencias de potencia del auto

Resulta que el 75% de los autos asegurados tienen potencia de hasta los 128 caballos de vapor.

3.4.2 ANÁLISIS DESCRIPTIVO BIVARIANTE.

Ahora se realiza el análisis descriptivo bivariante entre la variable dependiente, estado de la póliza y las variables explicativas. Se estudia la correlación que hay entre cancelar la póliza y las distintas variables explicativas, para ello se usa la correlación de *Spearman* denotada por la letra *ρ(rho),* la cual calcula la correlación entre el rango de una variable *x* y el rango de las otras variables y_j, cuya formula es la siguiente.

$$\rho = 1 - \frac{6 \cdot \sum_{i=1,j=1}^{n,k} d_{ij}^2}{n(n^2-1)}$$

Donde:

$y_i = estado\ de\ la\ poliza, \quad para\ i = 1 \dots n\ observaciones$
$x_{ij} = variables\ explicativas\ , \quad para\ j = 1 \dots k\ variables$
$n = \ numero\ de\ observaciones$
$d_{ij} = diferencia\ de\ rangos\ i\ de\ la\ variable\ k$

- **Variable sex_customer**

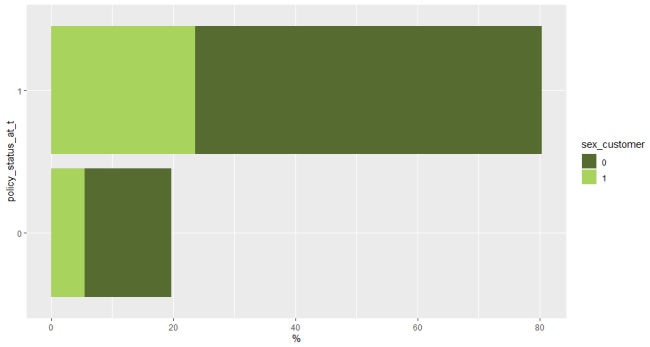


Figura 3.30 Diagrama de barras del estado de la poliza por sexo

Tabla cruzada variable dependiente: estado de la poliza

sex_customer	*policy_status_at_t* 0	1	***Total***
0	68936 19.9 %	276926 80.1 %	345862 100 %
1	26793 18.9 %	114955 81.1 %	141748 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=67.511 \cdot df=1 \cdot \varphi=0.012 \cdot p=0.000$

Tabla 3.27 Tabla cruzada estado de la póliza y sexo del cliente.

Como resumen de la información que presenta la tabla cruzada entre la variable dependiente estado de la póliza y el sexo del cliente, tanto de los 345.862 clientes masculinos como de los 141748 clientes de sexo femenino, el 19% canceló la póliza. Sin embargo, no hay correlación entre las variables.

- **Variable Policy_PaymentMethod**

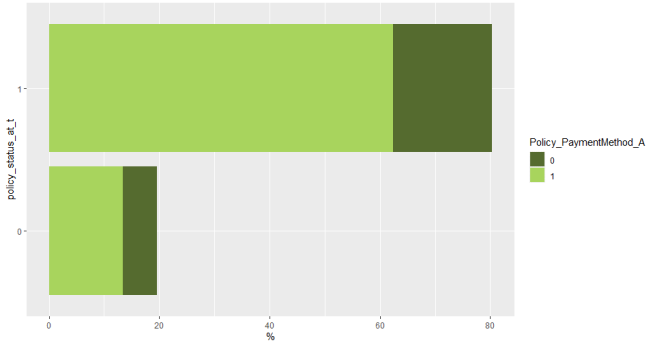


Figura 3.31 Diagrama de barras del estado de la póliza por método de pago

Tabla cruzada variable dependiente: estado de la poliza

Policy_PaymentMethod	*policy_status_at_t* 0	1	***Total***
0	30556 25.9 %	87468 74.1 %	118024 100 %
1	65173 17.6 %	304413 82.4 %	369586 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=3863.657 \cdot df=1 \cdot \varphi=0.089 \cdot p=0.000$

Tabla 3.28 Tabla cruzada estado de la póliza y el método de pago

Para la variable del método de pago se tiene que de las 369586 pólizas que pagaron de forma automática, el 17.6% canceló la póliza, mientras que de las 118024 que pagan a través de otros métodos el 26% canceló su póliza.

- **Variable totalpol**

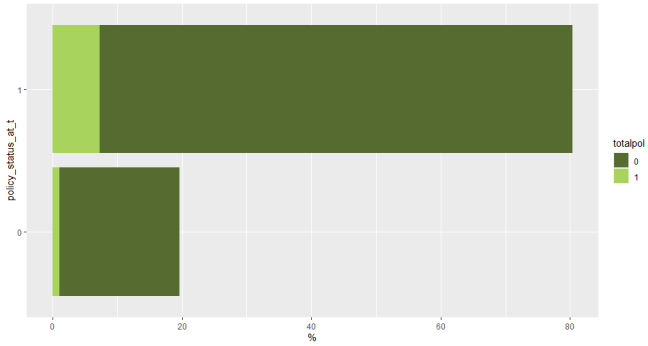


Figura 3.32 Diagrama de barras del estado de la póliza por poseer de pólizas adicionales

Tabla cruzada variable dependiente: estado de la poliza

totalpol	*policy_status_at_t* 0	1	***Total***
0	90857 20.3 %	356251 79.7 %	447108 100 %
1	4872 12 %	35630 88 %	40502 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=1617.864 \cdot df=1 \cdot \varphi=0.058 \cdot p=0.000$

Tabla 3.29 Tabla cruzada estado de la póliza y poseer pólizas adicionales

Para la variable de pólizas adicionales en resumen de la tabla *3.29*, de las 447108 pólizas que no tienen asociada otra póliza, el 20% canceló la póliza, mientras que el 12% de las 40502 que si tienen asociada otras pólizas canceló su póliza. Aun así, no existe correlación entre las variables.

- **Variable contracted_guarantee_M**

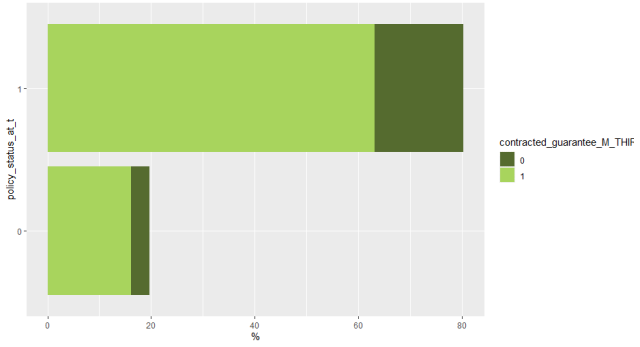


Figura 3.33 Diagrama de barras del estado de la póliza por la garantía

Tabla cruzada variable dependiente: estado de la poliza

contracted_guarantee	*policy_status_at_t* 0	1	***Total***
0	17082 16.9 %	84112 83.1 %	101194 100 %
1	78647 20.4 %	307769 79.6 %	386416 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=612.647 \cdot df=1 \cdot \varphi=0.035 \cdot p=0.000$

Tabla 3.30 Tabla cruzada estado de la póliza y la garantía contratada.

Como resumen de la información que presenta la tabla cruzada entre la variable dependiente estado de la póliza y la garantía contratada, de las 386416 pólizas contra terceros el 20% canceló su póliza, mientras que de los 101194 con garantías a todo riesgo con y sin franquicia, el 17% canceló su póliza. Aun así, no existe correlación entre las variables.

- **Variable car_bonus_M**

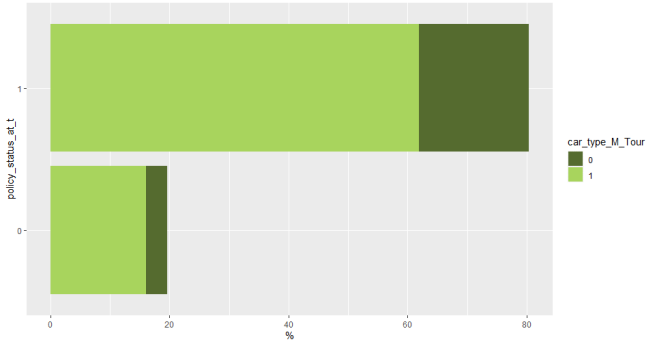


Figura 3.34 Diagrama de barras del estado de la póliza por poseer bono

Tabla cruzada variable dependiente: estado de la poliza

car_bonus_M	*policy_status_at_t* 0	1	***Total***
0	22506 20.9 %	85186 79.1 %	107692 100 %
1	73223 19.3 %	306695 80.7 %	379918 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=140.346 \cdot df=1 \cdot \varphi=0.017 \cdot p=0.000$

Tabla 3.31 Tabla cruzada estado de la póliza y el bono

Para la variable bono de las pólizas, de las 379918 pólizas que poseen un bono como de los 107692 que no poseen bono, sobre el 20% canceló la póliza.

- **Variable car_type_M**

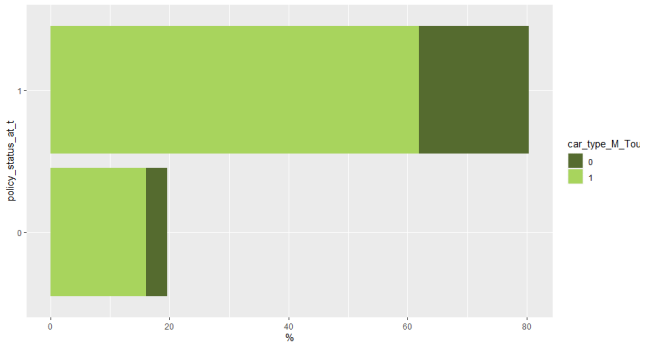


Figura 3.35 Diagrama de barras del estado de la póliza por el tipo de vehiculo

Tabla cruzada variable dependiente: estado de la poliza

car_type_M_Tour	*policy_status_at_t* 0	1	***Total***
0	17764 16.5 %	90173 83.5 %	107937 100 %
1	77965 20.5 %	301708 79.5 %	379673 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=885.147 \cdot df=1 \cdot \varphi=0.043 \cdot p=0.000$

Tabla 3.32 Tabla cruzada estado de la póliza y el tipo de auto

Como resumen la tabla cruzada entre la variable dependiente estado de la póliza y el tipo de auto, de las 379673 pólizas con autos de tipo turismo el 20.5 % canceló su póliza, mientras que de los 107937 con otro tipo de auto, el 16.5% canceló su póliza. Aun así, no existe correlación entre las variables.

- **Variable Car_2ndDriver_M**

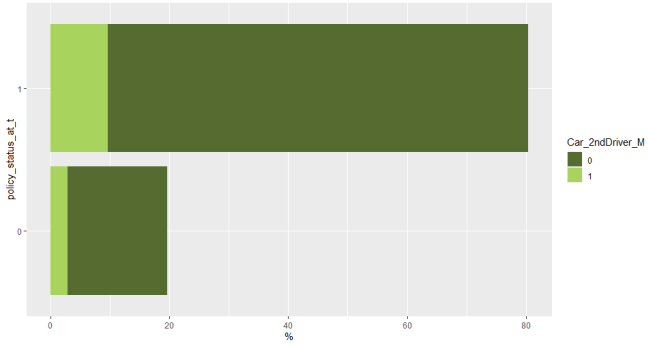


Figura 3.36 Diagrama de barras del estado de la póliza por poseer segundo conductor

Tabla cruzada variable dependiente: estado de la poliza

Car_2ndDriver_M	*policy_status_at_t* 0	1	***Total***
0	81790 19.2 %	344799 80.8 %	426589 100 %
1	13939 22.8 %	47082 77.2 %	61021 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=455.469 \cdot df=1 \cdot \varphi=0.031 \cdot p=0.000$

Tabla 3.33 Tabla cruzada estado de la póliza y segundo conductor

Para la variable de segundo conductor en resumen de la *tabla cruzada 3.33*, de las 61021 pólizas que tienen indicado a un segundo conductor, el 19% canceló la póliza, mientras que de las 426589 que no tienen indicado segundo conducto, el 23% canceló su póliza. Aun así, no existe correlación entre las variables.

- **Variable Fuel_Type_M**

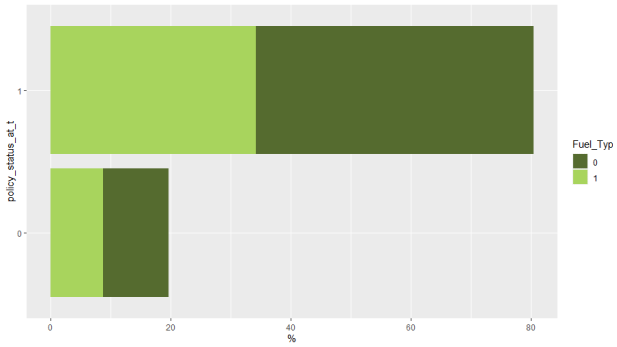


Figura 3.37 Diagrama de barras del estado de la póliza por el tipo de combustible

Tabla cruzada variable dependiente: estado de la poliza

Fuel_Type_M_G	*policy_status_at_t* 0	1	***Total***
0	53007 19 %	225308 81 %	278315 100 %
1	42722 20.4 %	166573 79.6 %	209295 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=141.325 \cdot df=1 \cdot \varphi=0.017 \cdot p=0.000$

Tabla 3.34 Tabla cruzada estado de la póliza y tipo de combustible

Como resumen de la información que presenta la tabla cruzada entre la variable dependiente estado de la póliza y el tipo de combustible, tanto para de las 209295 pólizas de vehículos de Diesel como de las 209295 pólizas de vehículos de gasolina, sobre el 19% canceló la póliza.

- **Variable Car_number_of_seats_M**

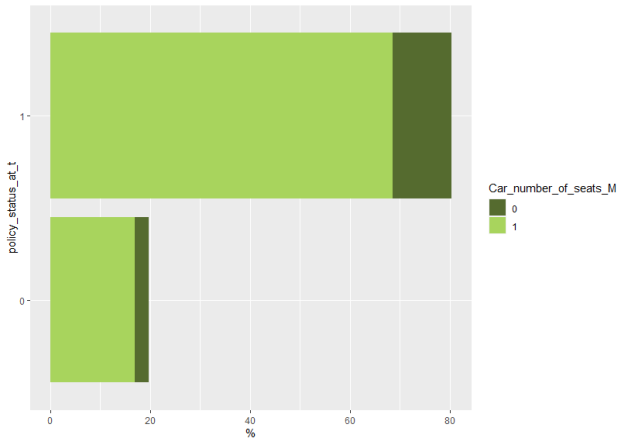


Figura 3.38 Diagrama de barras del estado de la póliza por el numero de asientos

Tabla cruzada variable dependiente: estado de la poliza

Car_number_of_seats	*policy_status_at_t* 0	1	***Total***
0	12941 18.4 %	57304 81.6 %	70245 100 %
1	82788 19.8 %	334577 80.2 %	417365 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=76.017 \cdot df=1 \cdot \varphi=0.012 \cdot p=0.000$

Tabla 3.35 Tabla cruzada estado de la póliza y número de asientos del auto

Para las 417365 pólizas con autos de 5 asientes y para las 70245 que poseen autos con 1,2,3,4,6,7,8 o 9 asientos el sobre 19% canceló su póliza.

- **Variable dif_current_previous**

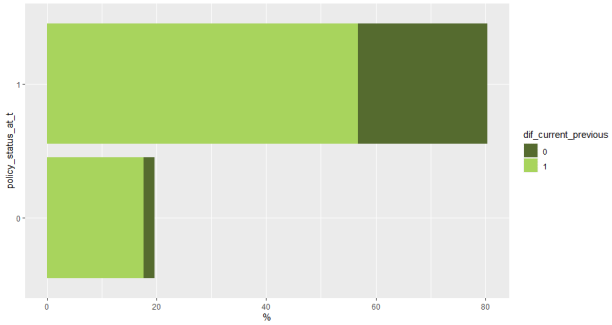


Figura 3.40 Diagrama de barras del estado de la póliza por la diferencia de la prima actual y previa

Tabla cruzada variable dependiente: estado de la poliza

dif_current_previous	*policy_status_at_t* 0	1	***Total***
0	9900 7.9 %	115233 92.1 %	125133 100 %
1	85829 23.7 %	276648 76.3 %	362477 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=14655.121 \cdot df=1 \cdot \varphi=0.173 \cdot p=0.000$

Tabla 3.36 Tabla cruzada estado de la póliza y diferencia prima actual y previa

En resumen, de las 125133 pólizas con diferencias negativas el 8% canceló su póliza mientras que para las 362477con diferencias positivas el 24% canceló su póliza. Aun así, no existe correlación entre las variables.

- **Variable dif_current_first**

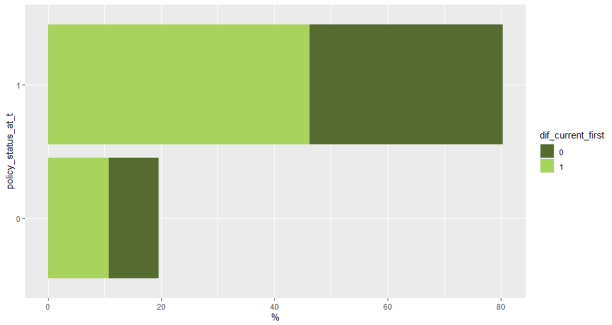


Figura 3.41 Diagrama de barras del estado de la póliza por la diferencia de la prima actual y primera

Tabla cruzada variable dependiente: estado de la poliza

dif_current_first	*policy_status_at_t* 0	1	***Total***
0	43408 20.7 %	166377 79.3 %	209785 100 %
1	52321 18.8 %	225504 81.2 %	277825 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=261.773 \cdot df=1 \cdot \varphi=0.023 \cdot p=0.000$

Tabla 3.37 Tabla cruzada estado de la póliza y diferencia prima actual y la primera

En resumen, de las 209785 pólizas con diferencias negativas y para las 277825 con diferencias positivas sobre el 19% canceló su póliza.

- **Variable Policy_numSupplements**

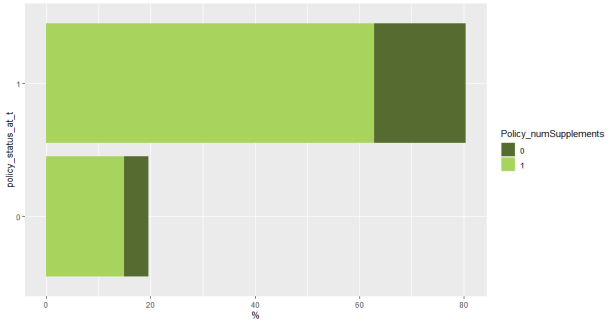


Figura 3.42 Diagrama de barras del estado de la póliza por poseer suplementos

Tabla cruzada variable dependiente: estado de la poliza

Policy_numSupplements	*policy_status_at_t* 0	1	***Total***
0	22506 20.9 %	85186 79.1 %	107692 100 %
1	73223 19.3 %	306695 80.7 %	379918 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=140.346 \cdot df=1 \cdot \varphi=0.017 \cdot p=0.000$

Tabla 3.38 Tabla cruzada estado de la póliza y poseer suplementos

Tanto las pólizas que poseen suplementos como las que no el 20 % canceló su póliza.

- **Variable Car_power_M**

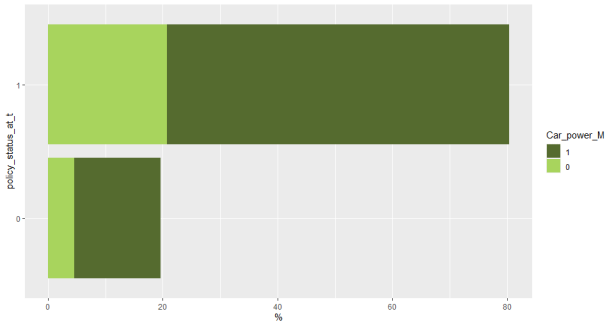


Figura 3.45 Diagrama de barras del estado de la póliza por la potencia del auto

Tabla cruzada variable dependiente: estado de la poliza

Car_power_M	*policy_status_at_t* 0	1	***Total***
1	73214 20.1 %	290897 79.9 %	364111 100 %
0	22515 18.2 %	100984 81.8 %	123499 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=205.734 \cdot df=1 \cdot \varphi=0.021 \cdot p=0.000$

Tabla 3.39 Tabla cruzada estado de la póliza y potencia del auto

Tanto para las pólizas que poseen autos con potencia menor de 128 CV, como las de más de 128CV, el sobre 19% canceló su póliza.

- **Variable value_of_car_9k**

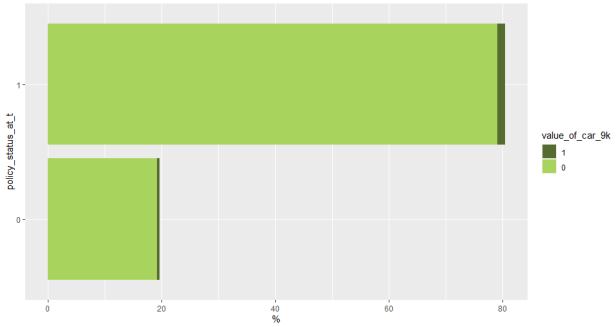


Figura 3.46 Diagrama de barras del estado de la póliza por valor de auto menor a 9k

Tabla cruzada variable dependiente: estado de la poliza

value_of_car_9k	*policy_status_at_t* 0	1	***Total***
1	2180 24.6 %	6693 75.4 %	8873 100 %
0	93549 19.5 %	385188 80.5 %	478737 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=139.272 \cdot df=1 \cdot \varphi=0.017 \cdot p=0.000$

Tabla 3.40 Tabla cruzada estado de la póliza y valor auto menor a 9k

En resumen, de las 8873 pólizas con autos con valores menores a €9000 el 25% canceló su póliza, mientras que para las 478737 con autos de superior valor el 20% canceló, sin embargo, no existe correlación entre las variables.

- **Variable value_of_car_M**

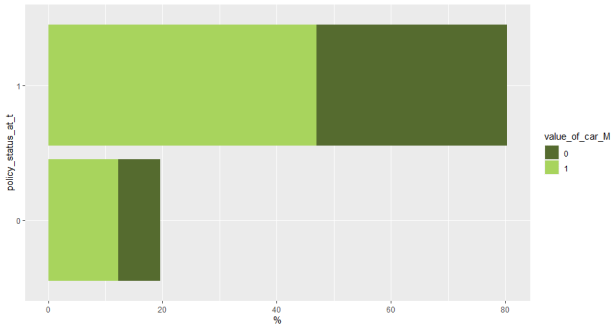


Figura 3.47 Diagrama de barras del estado de la póliza por valor de auto de 9k-22k

Tabla cruzada variable dependiente: estado de la poliza

value_of_car_M	*policy_status_at_t* 0	1	***Total***
0	36137 18.2 %	162802 81.8 %	198939 100 %
1	59592 20.6 %	229079 79.4 %	288671 100 %
Total	95729 19.6 %	391881 80.4 %	487610 100 %

$\chi^2=458.457 \cdot df=1 \cdot \varphi=0.031 \cdot p=0.000$

Tabla 3.41 Tabla cruzada estado de la póliza y valor auto de 9k-22k

Como resultado se obtuvo que tanto para autos asegurados entre €9000 y €22000 como para el resto de auto sobre el 19% cancela su póliza.

3.5 ANÁLISIS DESCRIPTIVO PÓLIZAS DE HOGAR

En este apartado se realiza un análisis univariante y bivariante equivalente al que se realizó para las pólizas de auto, pero, para las pólizas de hogar.

.

3.5.1 ANÁLISIS DESCRIPTIVO UNIVARIANTE.

De igual forma que para las pólizas de auto, se realiza el análisis univariante para las 6 variables numéricas y las 8 variables categóricas de las pólizas de hogar.

3.5.1.1 Variables cuantitativas.

- **Variable Age_client:**

  Esta variable numérica informa la edad del cliente poseedor de la póliza. Es una variable que toma valores enteros entre 18 y los 85 años.

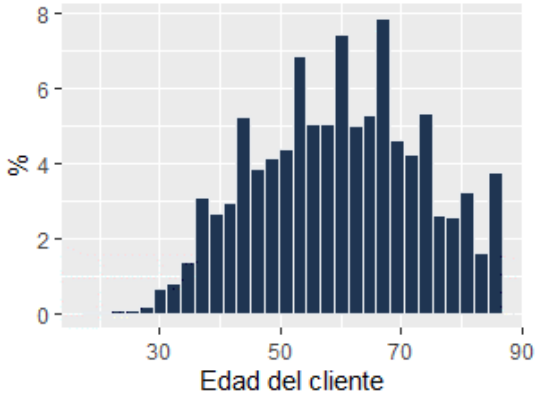


Figura 3.48 Histograma edad del cliente

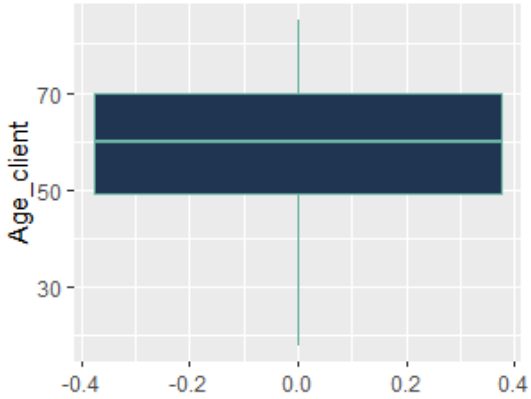


Figura 3.49 Boxplot edad del cliente

Min.	1stQu.	Median	Mean	3rdQu.	Max.
18.00	49.00	60.00	59.71	70.00	85.00

Tabla 3.42 Estadísticos de edad del cliente

Tanto en la figura como en la figura los datos parecen comportarse de forma normal con una tendencia hacia la derecha, es decir, hacia las edades superiores.

- **Variable Client_Seniority**

Esta variable permite conocer la antigüedad en años del cliente poseedor de la póliza en la compañía, toma valores entre 1 y 41 años. A continuación, se muestra una tabla de frecuencias y porcentajes para poder darnos una idea de esta antigüedad.

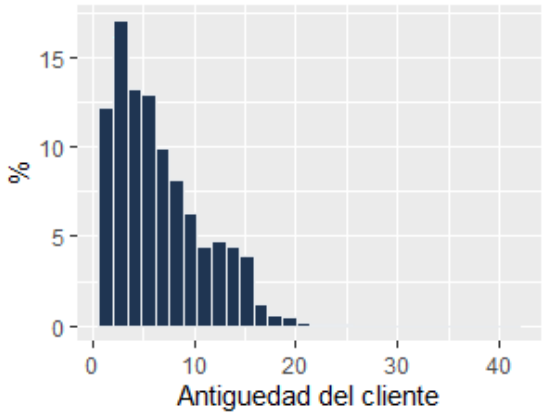


Figura 3.50 Histograma antigüedad del cliente

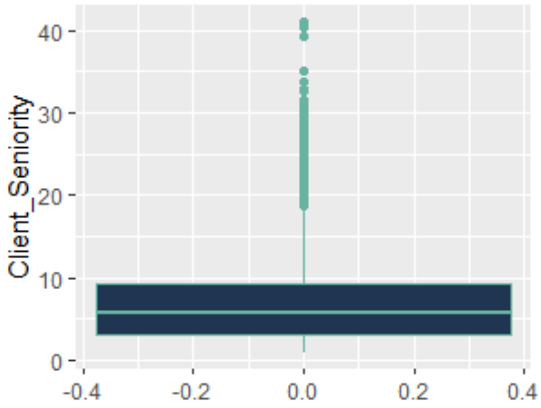


Figura 3.51 Boxplot antigüedad del cliente

Min.	1stQu.	Median	Mean	3rdQu.	Max.
1.002	3.072	5.645	6.672	9.311	41.051

Tabla 3.43 Estadísticos antigüedad del cliente

En promedio los clientes llevan 5,6 años con la póliza, aunque a partir del tercer cuartil de 9 años la antigüedad se dispersa hasta los 41 años.

- **Variable previous_to_last_premium_paid**

Esta variable representa la prima previa a la última prima paga, es una variable continua que toma valores desde 0 hasta los 5600.

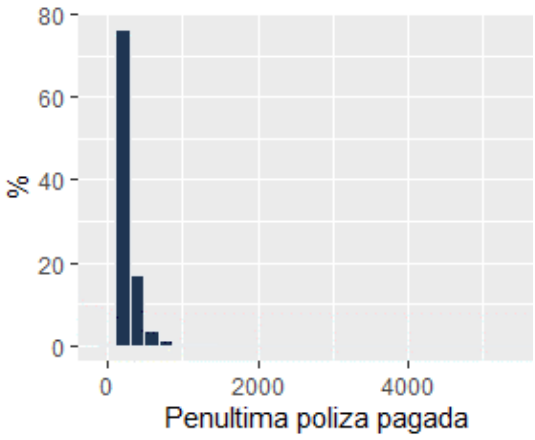


Figura 3.51 Histograma penúltima póliza pagada

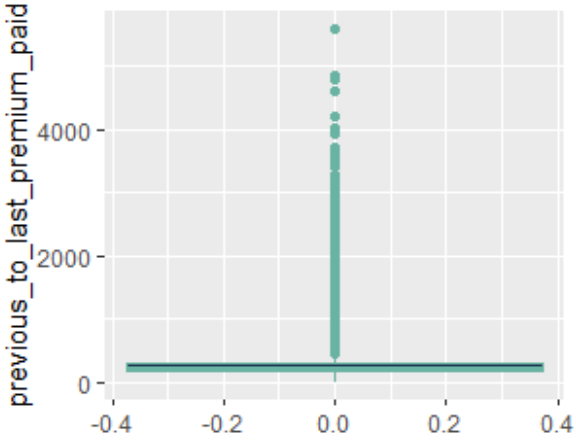


Figura 3.52 Boxplot penúltima póliza pagada

Min.	1stQu.	Median	Mean	3rdQu.	Max.
0.0	163.7	207.2	252.4	281.6	5599.8

Tabla 3.44 Estadísticos de penúltima póliza pagada

En promedio las primas valen €252 aunque existe una asimetría hacia la derecha en donde la prima puede llegar a tomar valores de hasta €5600.

- **Variable last_premium_paid**

La última prima pagada es una variable continua que toma valores desde 37 hasta 5600

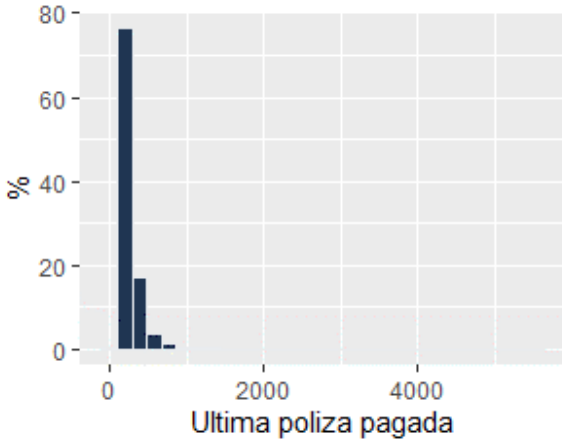


Figura 3.53 Histograma última póliza pagada

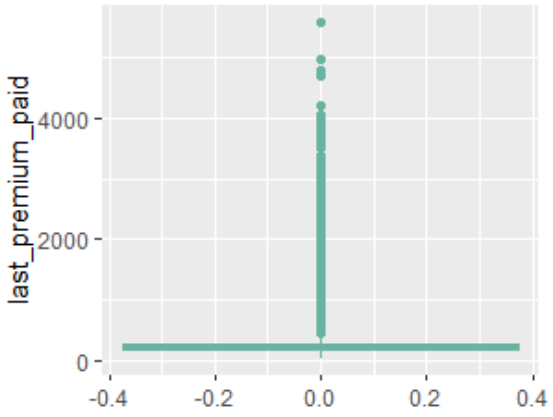


Figura 3.54 Boxplot última póliza pagada

Min.	1stQu.	Median	Mean	3rdQu.	Max.
37.58	161.31	204.44	249.79	278.46	5599.79

Tabla 3.45 Estadísticos última póliza pagada

En promedio las primas valen €250 aunque existe una asimetría hacia la derecha en donde la prima puede llegar a tomar valores de hasta €5600.

- **Variable Insuredcapital_content_H**

El contenido de capital asegurado es el monto máximo establecido que cubre la póliza. La variable es numérica continua y puede cubrir desde €1 hasta un poco más de un millón y medio (€1637499) de euros.

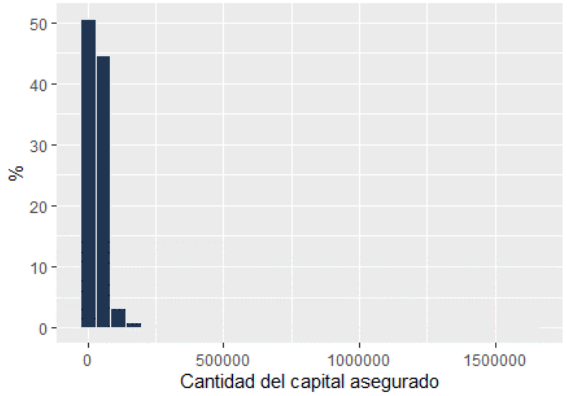


Figura 3.55 Histograma del capital asegurado

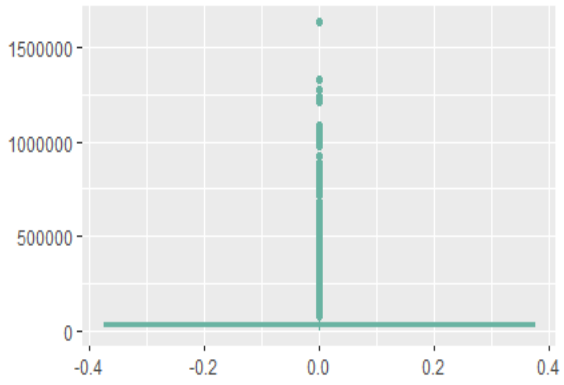


Figura 3.56 Boxplot del capital asegurado

Min.	1stQu.	Median	Mean	3rdQu.	Max.
1	18483	27956	34795	40637	1637499

Tabla 3.46 Estadísticos del capital asegurado

En promedio las pólizas cubren hasta los €34795 pero como se aprecia en el boxplot existe una gran dispersión,

- **Variable Insuredcapital_continent_H**

El continente del capital asegurado es el valor establecido del objeto en cuestión. Este valor es numérico y continuo y puede tomar valores desde los €180 hasta los casi 5 millones de euros.

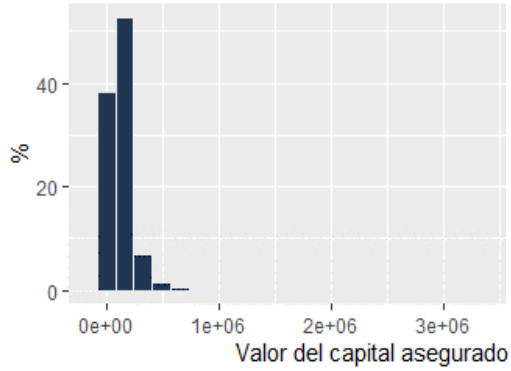


Figura 3.57 Histograma del valor del capital asegurado

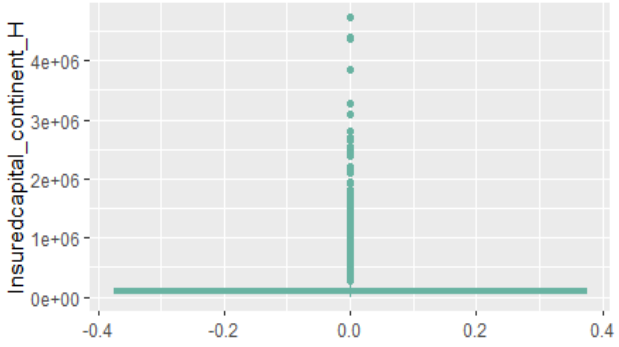


Figura 3.58 Boxplot del valor del capital asegurado

Min.	1stQu.	Median	Mean	3rdQu.	Max.
180	65782	96197	122804	150391	4740076

Tabla 3.47 Estadísticos del valor del capital asegurado

En promedios las propiedades aseguradas valen €22804 aunque igual se aprecia una gran dispersión existiendo propiedades de casi 5 millones de euros.

3.5.1.2 Variables categóricas

- **Variable sex_customer**

Esta variable informa del sexo del poseedor de la póliza, es binaria de forma que el 1 representa al sexo femenino y el 0 al masculino.

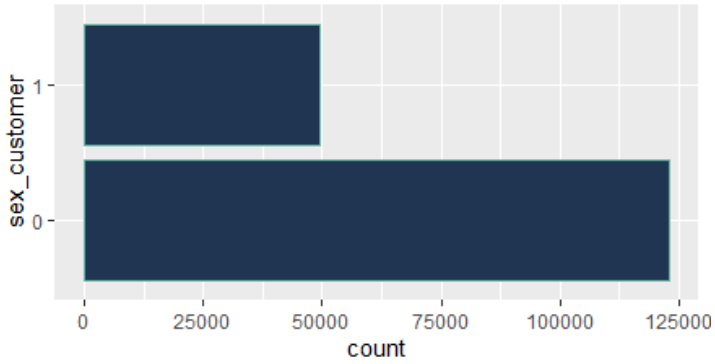


Figura 3.59 Diagrama de barras del sexo del cliente

sex_customer

	Frequency	Percent
0	123052	71.25
1	49656	28.75
Total	172708	100.00

Tabla 3.48 Frecuencia del sexo del cliente

Se puede apreciar en la tabla y gráficamente que el 71% son de sexo masculino y el 29% restante de sexo femenino.

- **Variable Policy_PaymentMethod**

Esta variable se transforma en binaria con el fin de crear clases más equilibradas, el 1 representa al pago automático y el 0 al resto de tipologías de pago, S, T y U.

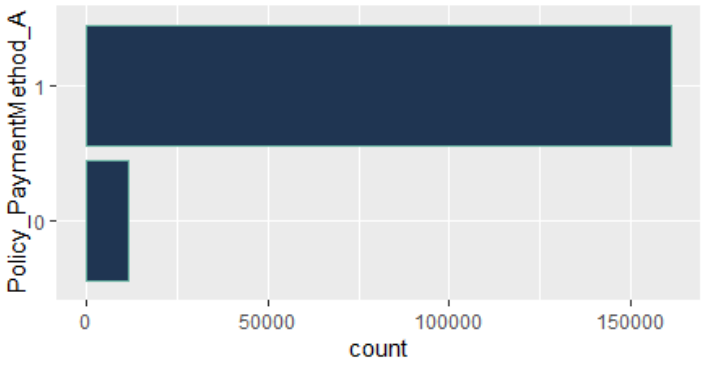


Figura 3.60 Diagrama de barras del método del pago

Policy_PaymentMethod_A

	Frequency	Percent
0	11521	6.671
1	161187	93.329
Total	172708	100.000

Tabla 3.49 Frecuencia del método de pago

El 93% de las pólizas son pagadas de forma automática, el 7% restante de forma S, T y U.

- **Variable policy_status_at_t**

El estado de la póliza es la variable dependiente del estudio de la fidelidad del cliente de este trabajo, indica el estado en el que se encuentra la póliza, es una variable binaria, el 1 representa a las pólizas vigentes y el 0 a las anuladas.

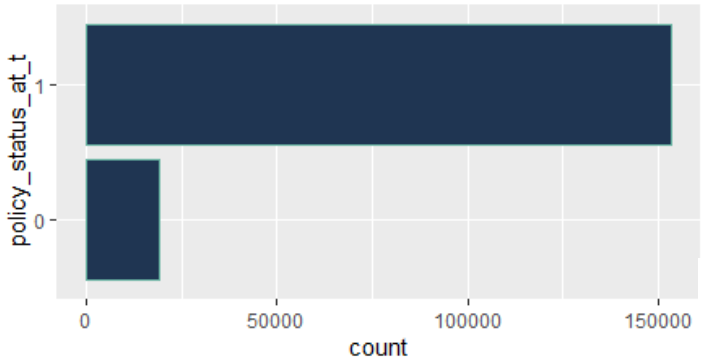


Figura 3.61 Diagrama de barras del estado de la póliza

policy_status_at_t

	Frequency	Percent
0	19137	11.08
1	153571	88.92
Total	172708	100.00

Tabla 3.50 Frecuencia de barras del estado de la póliza

Se evidencia la desproporción entre las dos clases, las pólizas vigentes representan un 89% de todas las pólizas de hogar.

- **Variable pclaims**

La probabilidad de declarar siniestro es la otra variable binaria de estudio, el 0 representa 0 siniestros declarados y el 1 representa que ha sufrido algún siniestro.

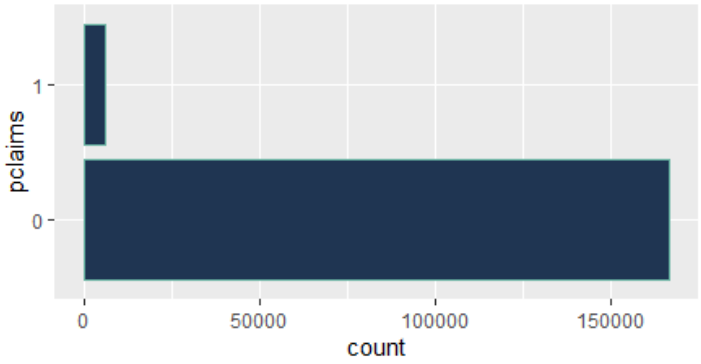


Figura 3.62 Diagrama de barras de reclamaciones

pclaims

	Frequency	Percent
0	166618	96.474
1	6090	3.526
Total	172708	100.000

Tabla 3.51 Frecuencia de barras de las reclamaciones

La probabilidad de declarar siniestro en las pólizas del hogar es de 8,5%

- **Variable totalpol**

La variable total de pólizas es una variable binaria e indica si el cliente poseedor de la póliza tiene asociada otras pólizas o no, el 0 representa que no es poseedor de ninguna otra póliza y el 1 que al menos posee una póliza más.

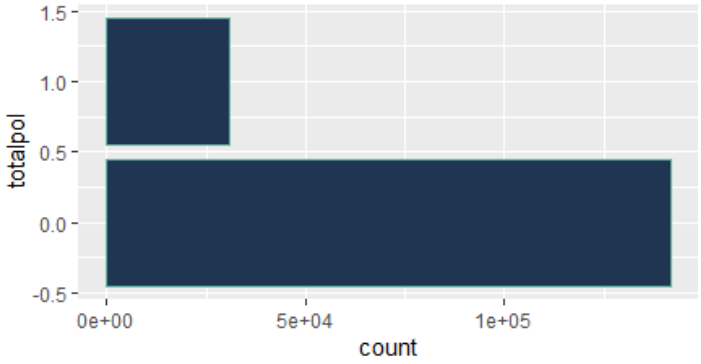


Figura 3.63 Diagrama de barras de las pólizas adicionales

totalpol

	Frequency	Percent
0	141802	82.11
1	30906	17.89
Total	172708	100.00

Tabla 3.52 Frecuencia de las pólizas adicionales

El 82% de las pólizas no posee otra póliza asociada y el 18% si tiene asociadas alguna otra póliza.

- **dif_current_previous**

La variable de la diferencia entre la prima previa y la actual indica 1 si la diferencia es positiva y 0 si es negativa.

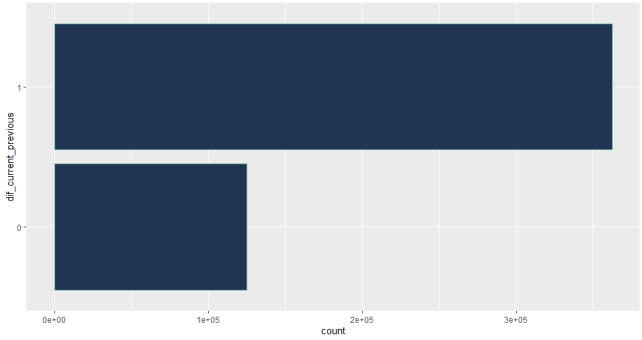


Figura 3.64 Diagrama de barras diferencias

totalpol

	Frequency	Percent
0	25452	14.74
1	147256	85.26
Total	172708	100.000

Tabla 3.53 Frecuencias de pólizas adicionales

Se observa que el 15% de las pólizas tuvieron una diferencia negativa significando que sus primas actuales disminuyeron con respecto a la anterior.

- **dif_current_first**

La diferencia entre la póliza actual y la primera es una variable binaria que indica 1 si la diferencia es positiva y 0 si es negativa.

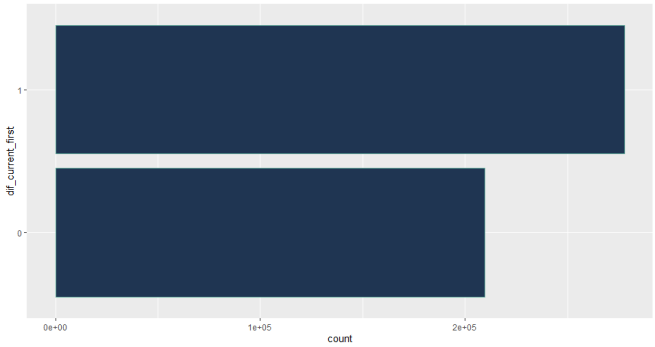


Figura 3.65 Diagrama de barras diferencias

totalpol

	Frequency	Percent
0	9713	5.62
1	162995	94.38
Total	172708	100.000

Tabla 3.54 Frecuencias de pólizas adicionales

Resulta que el 6% de las pólizas tuvieron primas previas que disminuyeron respecto a la primera prima que pagó.

- **Variable Home_Type**

Las pólizas del hogar se pueden clasificar según el tipo de hogar asegurado, esta variable se transformó en binaria e indica 1 si el tipo de hogar es un PI o 0 si es AT, PB, RU, UA y UF

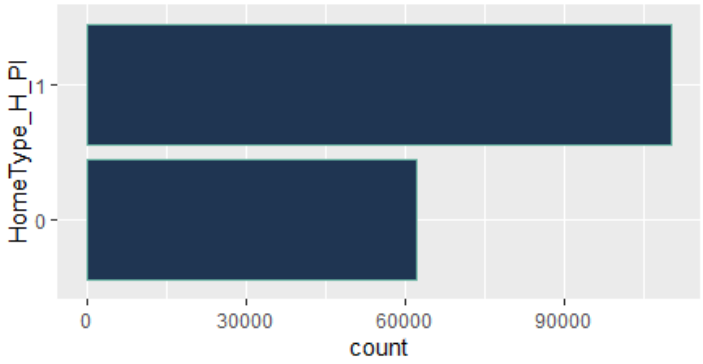


Figura 3.66 Diagrama de barras del tipo de hogar

HomeType_H_Pl

	Frequency	Percent
0	62296	36.07
1	110412	63.93
Total	172708	100.00

Tabla 3.55 Frecuencia del tipo de hogar

El 64% de las pólizas de hogar poseen un hogar tipo piso, el 36% restante posee otro tipo de patrimonio.

3.5.2 ANÁLISIS DESCRIPTIVO BIVARIANTE.

A continuación, análisis descriptivo bivariante entre la variable dependiente estado de la póliza y las variables explicativas para las pólizas de hogar.

- **Variable sex_customer**

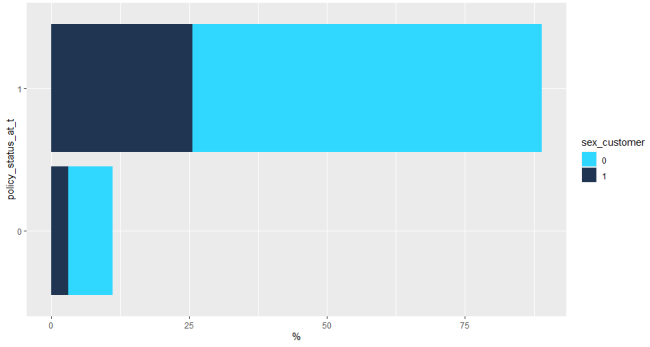


Figura 3.67 Diagrama de barras del estado de la póliza por el sexo del cliente

Tabla cruzada variable dependiente: estado de la poliza

sex_customer	*policy_status_at_t* 0	1	***Total***
0	13710 11.1 %	109342 88.9 %	123052 100 %
1	5427 10.9 %	44229 89.1 %	49656 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=1.599 \cdot df=1 \cdot \varphi=0.003 \cdot p=0.206$

Tabla 3.56 Tabla cruzada estado de la póliza y sexo del cliente

Como resumen de la información que presenta la tabla cruzada entre la variable dependiente estado de la póliza y el sexo del cliente, tanto de los 123052 clientes masculinos como de los

49656 clientes de sexo femenino, el 11% canceló la póliza. Asimismo, resulta haber una correlación de 0.21 entre el sexo y el estado de la póliza.

- **Variable Policy_PaymentMethod**

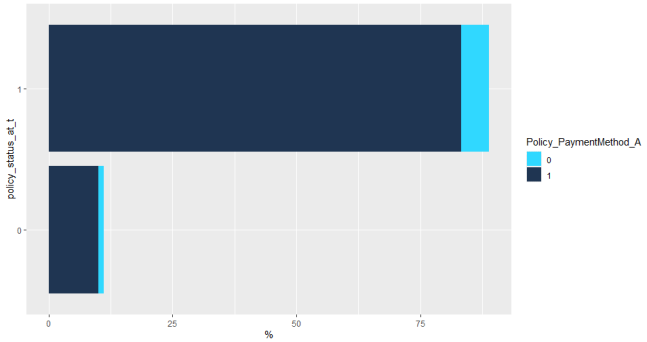


Figura 3.68 Diagrama de barras del estado de la póliza por el método de pago

Tabla cruzada variable dependiente: estado de la poliza

Policy_PaymentMethod	*policy_status_at_t* 0	1	***Total***
0	1759 15.3 %	9762 84.7 %	11521 100 %
1	17378 10.8 %	143809 89.2 %	161187 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=219.212 \cdot df=1 \cdot \varphi=0.036 \cdot p=0.000$

Tabla 3.57 Tabla cruzada estado de la póliza y método de pago

Para la variable del método de pago se tiene que de las 161187 pólizas que pagaron de forma anual, el 15% canceló la póliza, mientras que de las 11521 que pagan a través de otros métodos el 11% canceló su póliza, sin embargo, no hay correlación entre las variables.

- **Variable totalpol**

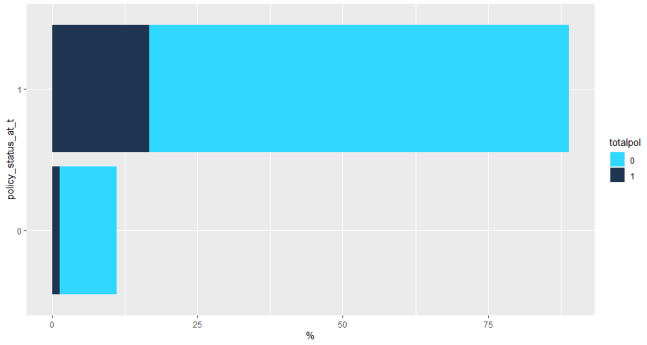


Figura 3.69 Diagrama de barras del estado de la póliza poseer pólizas adicionales

Tabla cruzada variable dependiente: estado de la poliza

totalpol	*policy_status_at_t* 0	1	***Total***
0	16992 12 %	124810 88 %	141802 100 %
1	2145 6.9 %	28761 93.1 %	30906 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=654.347 \cdot df=1 \cdot \varphi=0.062 \cdot p=0.000$

Tabla 3.58 Tabla cruzada estado de la póliza y poseer pólizas adicionales

Para la variable de pólizas adicionales en resumen de la *tabla 3.58*, de las 141802 pólizas que no tienen asociada otra póliza, el 12% la canceló, mientras que de las 30906 que, si tienen asociada otras pólizas, el 7% la canceló, pero, esta diferencia no conlleva una correlación entre las variables.

- **Variable HomeType_H**

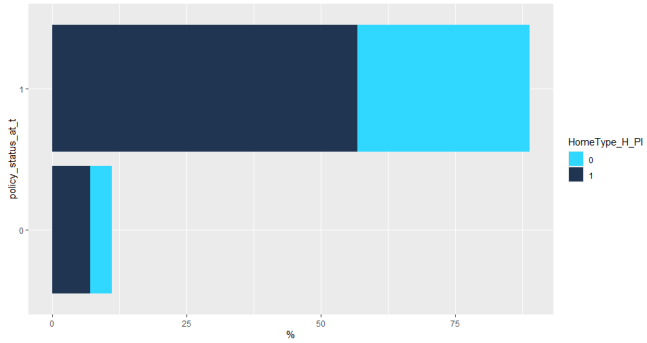


Figura 3.70 Diagrama de barras del estado de la póliza por el tipo de hogar

Tabla cruzada variable dependiente: estado de la poliza

HomeType_H_PI	*policy_status_at_t* 0	1	***Total***
0	6928 11.1 %	55368 88.9 %	62296 100 %
1	12209 11.1 %	98203 88.9 %	110412 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=0.156 \cdot df=1 \cdot \varphi=0.001 \cdot p=0.693$

Tabla 3.59 Tabla cruzada estado de la póliza y tipo de hogar

Como resumen de la información que presenta la tabla cruzada entre la variable dependiente estado de la póliza y el tipo de hogar, tanto de los 110412 de las pólizas de hogares de tipo piso como de los 62296 con otro tipo de hogar, el 11% canceló la póliza.

- **Variable dif_current_previous**

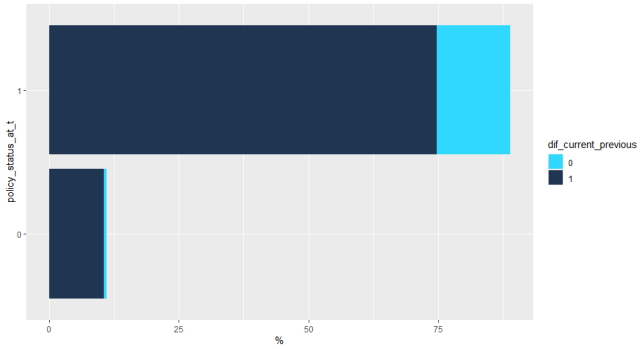


Figura 3.71 Diagrama de barras del estado de la póliza por la diferencia prima actual y previa

Tabla cruzada variable dependiente: estado de la poliza

dif_current_previous	*policy_status_at_t* 0	1	***Total***
0	952 3.7 %	24500 96.3 %	25452 100 %
1	18185 12.3 %	129071 87.7 %	147256 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=1631.488 \cdot df=1 \cdot \varphi=0.097 \cdot p=0.000$

Tabla 3.60 Tabla cruzada estado de la póliza y diferencia prima actual y previa

Para las 25452 diferencias negativas el 3.7% canceló su póliza, mientras que para el 147256 el 12% canceló su póliza, no aparenta haber correlacion entre las diferencias y el estado de la póliza.

- **Variable dif_current_first**

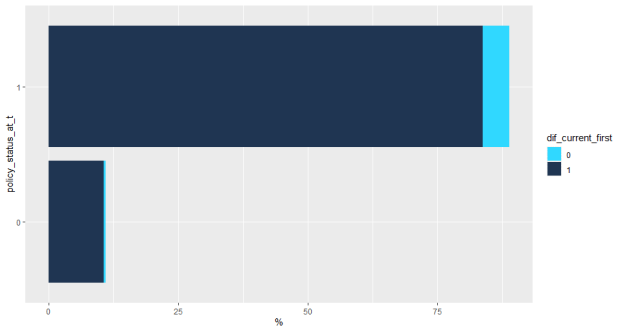


Figura 3.72 Diagrama de barras del estado de la póliza por la diferencia prima actual y primera

Tabla cruzada variable dependiente: estado de la poliza

dif_current_first	*policy_status_at_t* 0	1	***Total***
0	722 7.4 %	8991 92.6 %	9713 100 %
1	18415 11.3 %	144580 88.7 %	162995 100 %
Total	19137 11.1 %	153571 88.9 %	172708 100 %

$\chi^2=138.557 \cdot df=1 \cdot \varphi=0.028 \cdot p=0.000$

Tabla 3.61 Tabla cruzada estado de la póliza y diferencia prima actual y primera

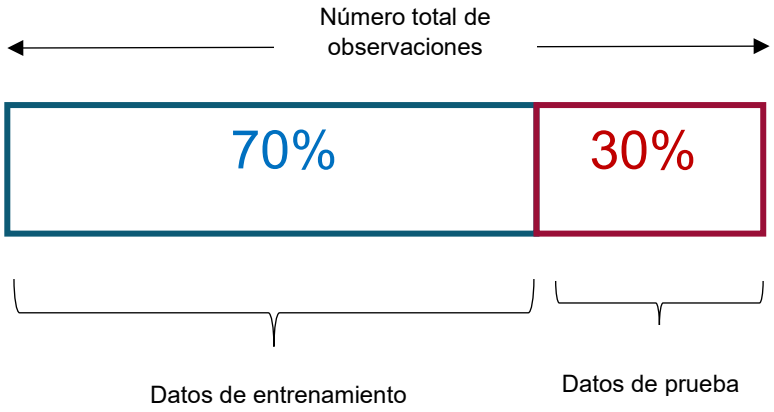
En resumen, de las 9713 pólizas con diferencia de primas negativas el 7% las canceló, mientras que de las 162995 con la diferencia positiva el 11% canceló, aunque existe esta diferencia de porcentajes no aparente haber correlación entre las variables.

3.6 SELECCIÓN DE LAS BASES DE DATOS DE ENTRENAMIENTO Y PRUEBA

Los datos de entrenamiento (training) y prueba (test) se usan generalmente en el aprendizaje supervisado de *machine learning* para evaluar la capacidad predictiva de los algoritmos.

La base de datos de entrenamiento, como su nombre lo indica se usa para enseñar al método a reconocer los patrones propios de los datos, lo que puede tener distintos usos, como por ejemplo, predecir la variable dependiente. En cambio, el conjunto de datos de prueba es un conjunto de datos que se utiliza para proporcionar una evaluación imparcial de un ajuste final del modelo resultante del conjunto de datos de entrenamiento.

La forma de selección de los datos se suele hacer dividiendo los datos originales en proporciones 70% de los datos para el entrenamiento y 30% para probar dicho entrenamiento, éstas proporciones pueden variar en función del número de observaciones disponibles.



Los datos al proceder de una sola matriz, el modelo resultante no podrá reconocer otros patrones por lo que no asegura que se pueda utilizar en una matriz de datos distinta. Para este problema existen experimentos, que han demostrado que el entrenamiento con datos auxiliares produce grandes mejoras en la precisión.

Se comprueba a través de las tablas del anexo 9.13.6 que efectivamente los datos de entrenamiento y prueba son representativos de las observaciones con una variación del ± 0.2%.

IV. METODOS DE MACHINE LEARNING

4.1 ASPECTOS GENERALES

Se utilizaron técnicas de *machine learning* por su capacidad de procesar modelos para clasificar grandes flujos de datos, el *machine learning* es un subconjunto de la inteligencia artificial que se

basa en programar una máquina para hacer inferencias generalizables sobre algún tipo de datos basados en datos anteriores.

En particular, definimos el machine learning como un conjunto de métodos que pueden detectar automáticamente patrones en los datos y luego usa los patrones descubiertos para predecir el dato futuro, o para realizar otros tipos de toma de decisiones en condiciones de incertidumbre[1]

Existen distintos métodos de aprendizaje, el supervisado es donde la salida del algoritmo se programa reconociendo los patrones con una salida objetivo, proporcionada por los datos observados (patrón), en cambio, el aprendizaje sin supervisión, no existen patrones objetivo, el algoritmo reconoce características en el conjunto de entradas que permiten formar grupos por similitudes.

La teoría de la probabilidad se puede aplicar a cualquier problema que implique incertidumbre, en los modelos estadísticos se basan en espacios de probabilidad *(Ω, F, P),* donde *Ω* es espacio muestral, *F* es una familia de subconjuntos y *P* es la función de probabilidad. Los algoritmos del *machine learning* también se basan en la teoría de probabilidad estadística y su noción axiomática de espacios de probabilidad, pero se usa un conjunto de datos, que se denota como:

$$S = \{(x_{ij}, y_{ij})\}, \quad donde$$

$x_{ij} = variables\ explicativas,\ para\ i = 1 \dots n\ observaciones\ y\ j = 1 \dots k\ variables$
$y_{ij} = variable\ respuesta\ para\ la\ i\ ésima\ observación\ de\ la\ variable\ k$

Este conjunto de datos de *n* observaciones y k variables se describe por características, que son proporcionados por *x,* y estas características están mapeadas por una determinada función para dar el valor *y.*

Para modelizar la probabilidad de cancelar la póliza, se usan los modelos de aprendizaje supervisado de **NN-Neural network** y de **SVM support vector machine** cuya principal difrerencia es ,que las NN dismunuyen el error que depende de la diferencia entre estos patrones de entrada y salida y las SVM categoriza de forma que maximiza la distancia del hiperplano que separa las clases.

4.1.1 Dificultades del Machine learning

La aplicación de la metodología de *machine learning* tiene como cualquier otro método sus particularidades que se podrían llamar dificultades o retos, que hay que tener en cuenta antes de aplicarlo, el sobreajuste, selección de datos, la predicción y las clases no balanceadas son los que se abordaron.

[1] MURPHY, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012. Pag 1

4.1.1.1 Sobreajuste

El *machine learning* no paramétrico al no ser aplicables a otros datos por su forma de entrenarse conlleva a un **sobreajuste**. Para prevenir este riesgo de sobreajuste, se utiliza la estrategia anteriormente mencionada de dividir los datos en **entrenamiento** (training data) para entrenar los modelos y **prueba** (test data) para evaluar el ajuste.

4.1.1.2 Relevancia de los datos

Los datos utilizados para entrenar los modelos predictivos han de ser de calidad y una cantidad significativa en el caso del aprendizaje automático. Los conjuntos de datos deben ser representativos y equilibrados para que puedan entrenar un modelo lo más optimo posible y evitar sesgos. Por eso se realizó el estudio exhaustivo de la matriz de datos en el apartado anterior.

4.1.1.3 Dificultad para predecir los rendimientos

Para valorar la capacidad predictiva de los modelos binarios se utilizó la **matriz de confusión**. La cual consiste en una tabla de doble entrada (puede tener otra dimensión) que por filas se encuentran los valores predichos y por columnas las observaciones para cada caso.

En la tabla 4.1 se muestra un ejemplo de una matriz binaria, que consta de 4 medidas:

- Verdaderos negativos: Datos observados negativos que el modelo clasifico como negativo.
- Falso positivos: Datos observados negativos que el modelo clasifico como positivos.
- Falso negativos: Datos observados positivos que el modelo clasifico como negativo.
- Verdaderos positivos: Datos observados positivos que el modelo clasifico como positivos.

	REAL=0 (P)	**REAL=1 (N)**
Predicción = 0	TN: Verdadero positivo	FN: Falso positivo
Predicción = 1	FP: Falso negativo	TP: Verdadero negativo

Tabla 4.1 Matriz de confusión para evaluar la capacidad predictiva

Las medidas de la matriz de confusión pueden usarse para evaluar la capacidad predictiva de varias componentes de un estudio:

- Precisión: Proporción de positivos pronosticados entre todas las observaciones.

$$Accuracy = \frac{(TP + TN)}{(P + N)}$$

- Exactitud: Proporción de positivos observados resultados entre los pronosticados como positivos.

$$Exactirud = \frac{TP}{(TP + FP)}$$

- Sensibilidad: Proporción de pronosticados como positivos.

$$Sens = \frac{TP}{P}$$

- Especificidad: Proporción de predicho como negativo.

$$SP = \frac{TN}{N}$$

Gráficamente se puede evaluar por medio del análisis de curvas ROC (*Receiver Operating Characteristic*), esta representa la tasa de verdaderos positivos (Sens) frente a la tasa de falsos negativos (1-Sp), por lo que describe y compara la precisión de las predicciones.

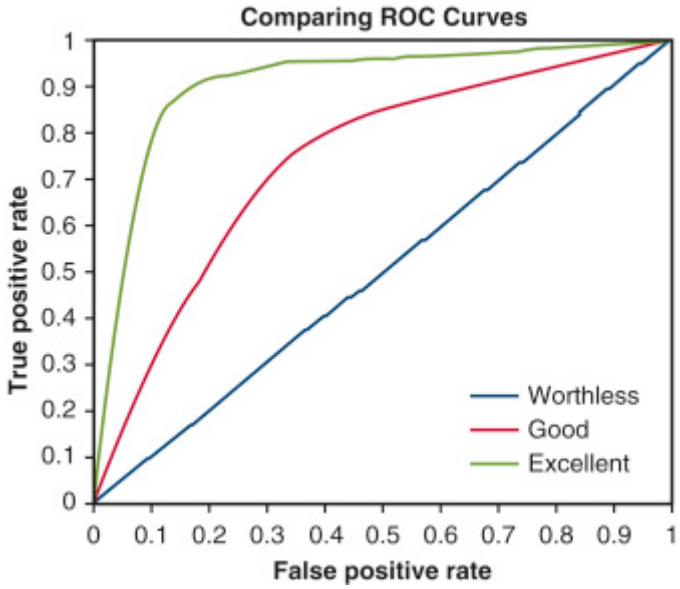


Figura 4.1 Interpretación curva ROC

https://marlin-prod.literatumonline.com/cms/attachment/36cdb4ec-0c7d-48cb-9a4d-7cb463f8b7c3/gr1.jpg

Si la curva ROC aumenta rápidamente hacia la esquina superior izquierda del gráfico, o si el valor del área bajo la curva (AUC) es > 0.5, el modelo predice de forma representativa a las observaciones. Por otro lado, si el AUC es ≤ 0.5 la predicción es insignificante, ya que, es similar a una selección aleatoria.

4.1.1.4 Datos no balanceados

Los datos no balanceados en la variable respuesta afecta a los algoritmos en su proceso de generalización de la información y perjudica a la clase minoritaria. Esto es debido a que, si los valores de entrada son 90% de una clase y solo el 10% de la otra, no diferenciará entre una clase de otra, lo más probable que el algoritmo se limite a responder siempre con la clase mayoritaria.

4.2 NEURAL NETWORK (NN)

Las redes neuronales artificiales conocidas por *NN-neural network* son estructuras de clasificación de información cuyo sistema y funcionamiento, están inspirados en las redes neuronales biológicas, por eso a los elementos de la red se llaman neuronas, estas neuronas o nodos están organizados en capas, a su vez cada neurona está conectada con otras neuronas mediante enlaces de comunicación y cada enlace tiene asociado un peso. Las principales características son:

1. *Autoorganización y Adaptabilidad: utilizan algoritmos de aprendizaje adaptativo y autoorganización, por lo que ofrecen mejores posibilidades de procesado robusto y adaptativo.*

2. *Procesado no Lineal: aumenta la capacidad de la red para aproximar funciones, clasificar patrones y aumenta su inmunidad frente al ruido.*

3. *Procesado Paralelo: normalmente se usa un gran número de nodos de procesado, con alto nivel de interconectividad.* [2]

Algunos aspectos generales de las NN en el área actuarial es que no sustituye los métodos lineales generalizados. La mayor diferencia entre los modelos, es que una NN busca minimizar el error entre el valor predicho y el observado y un modelo GLM, se basan en la maximización de la función de verosimilitud. Por eso las NN se usan como apoyo y generalmente se utiliza de una sola capa, ya que, son suficientes para predecir.

4.2.1 Tipos

Las NN supervisadas pueden se pueden clasificar principalmente por 3 factores:

1. Tipología: el tipo de red neuronal indica si las neuronas de una de las capas de la red pueden estar conectadas entre sí.

[2] Juan Miguel Marín Diazaraque, “Introducción a las redes neuronales aplicadas”, pp. 12 (2012)

- *Feedforward*: La retroalimentación permite conexiones de nodos entre dos capas diferentes
- *Feedback*: Permiten conexiones de nodos entre dos capas diferentes y conexiones entre de nodos de la misma capa.

2. Algoritmo de aprendizaje:

    - *Forwardpropagation*: Los valores de salida de los nodos de una NN solo se propagan a través de la red en una dirección, desde la capa de entrada a la capa de salida.
    - *Backpropagation*: Tiene una capa de entrada, una capa de salida y al menos una capa oculta. Se utiliza principalmente para la asociación de patrones.
    - *Selforganization*: Durante su proceso de aprendizaje, los nodos del mapa de características de la red se organizan en función de los valores de entrada dados.

3. Función de transferencia: Es la función aplicada a la combinación línea de las variables con sus respectivos pesos para dar el valor predicho de x, pueden ser de tipo:

    - Identidad
    - Escalón
    - Lineal a tramos
    - Gaussiana
    - Sigmoidal
    - Sinusoidal

4.2.2 Estructura

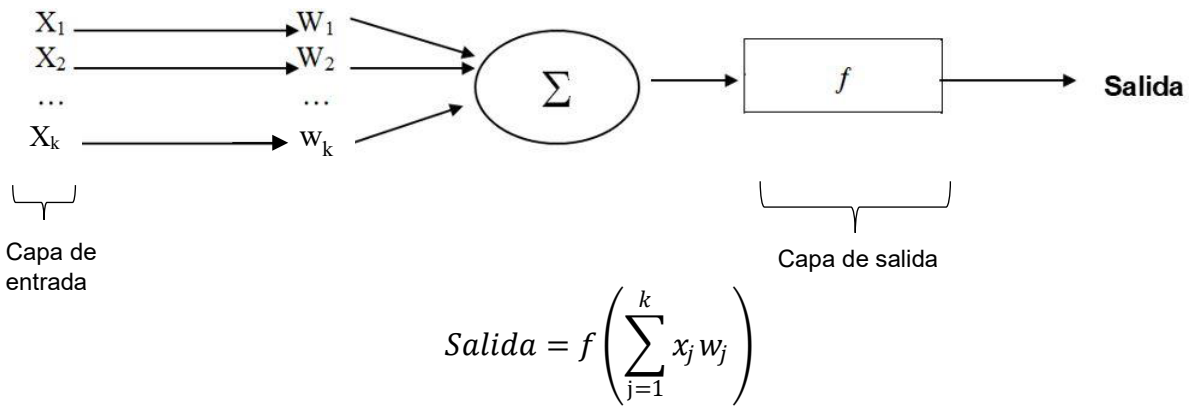
Las NN constan de 3 capas:

- Capa de entrada: Nodos que proporcionan los patrones de entrada en la red.

- Capas ocultas: Nodos cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.

- Capa de salida: Nodo que corresponde con la salida de la red, en algunos casos puede haber más de uno.

Gráficamente se tendría la siguiente estructura:



$$Salida = f\left(\sum_{j=1}^{k} x_j w_j\right)$$

Figura 4.2. Perceptrón de una red neuronal.

Donde:
$x_{11} \dots x_{ij} = valor\ de\ la\ variable\ j\ en\ el\ individuo\ i, \quad para\ i = 1 \dots n\ y\ j = 1 \dots k$
$w_{11} \dots w_{ij} = pesos\ asociados\ a\ las\ x_{ij}\ , \quad para\ i = 1 \dots n\ individuo\ y\ j = 1 \dots k\ variables$
$f = función\ de\ tranferencia$

Finalmente, para obtener los datos de salida se aplica la función de transferencia a la combinación lineal del valor de los nodos (k variables) con sus pesos correspondientes, como se muestra en la figura 4.2.

4.2.3 Funcionamiento

En resumen, lo que se desea para modelizar la probabilidad de que el asegurado cancele la póliza es entrenar una NN supervisada retroalimentada y de con algoritmo *backpropagation*, de la siguiente forma:

1. Asignarle unos **pesos** mediante enlaces de comunicación a la serie de datos observados que corresponden a las variables explicativas. Como explica *Diazaraque* (2012) «Dado un nuevo patrón de entrenamiento, en la etapa (m+1)-ésima, los pesos se adaptan de la siguiente forma:

$$w_{ij}^{m+1} = w_{ij}^{m} + \Delta w_{ij}^{m}, \quad para\ i = 1 \dots n\ y\ j = 1 \dots k$$

2. A cada nodo (neurona) se le asocia un umbral θ, por eso la importancia de que las variables se transformen a binaria. La operación efectuada por el perceptrón simple consiste en:

$$y = \begin{cases} 1\ si \sum_{j=1}^{k} x_j w_j \geq \theta\ , & para\ j = 1 \dots k \\ 0\ si \sum_{j=1}^{k} x_j w_j < \theta\ , & para\ j = 1 \dots k \end{cases}$$

3. Realizar una fase de **propagación hacia adelante** (*Forwardpropagation*). De forma que se predicen a través de la función los datos de salida esperados que ya fueron definidos por la variable estado de la póliza

4. Conseguir el **menor error** de salida para cambiar los valores de los pesos en dirección hacia atrás (***Backpropagation***).

5. Aplicarle la **función de transferencia sigmoidal** a la combinación lineal de los pesos y valores de las variables de la siguiente forma:

$$f\ salida = \frac{1}{1 + e^{-\Sigma x_j w_j}}\ , \qquad para\ \ j = 1 \dots k$$

6. El algoritmo finaliza, si todos los patrones de salida coinciden con sus patrones de destino.

4.3 SVM SUPPORT VECTOR MACHINE

El *SVM support vector machine* o máquina de soporte vectorial es otro método de *machine learning* que se puede utilizar para predecir la probabilidad de que el asegurado cancele su póliza. También es el caso de un algoritmo de aprendizaje supervisado por lo que requieren unos datos explicativos para una variable dependiente los cuales categoriza por medio de construir un hiperplano o conjunto de hiperplanos en un espacio.

Un hiperplano es subespacio plano de una (o varias) dimensiones, es decir, es una línea que divide las clases y la idea es encontrar el máximo margen de espacio para esta línea, así las clases estarían lo más separadas posibles lo que supondrá una definición más clara para cada clase y mejor predicción, matemáticamente un hiperplano de dimensión *p* para observaciones linealmente separables de clase 0 y 1, se tendría:

$$Hiperplano = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0,\ si\ \ y_i = 0 \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\ > 0,\ si\ \ y_i = \ 1 \end{cases}$$

Donde:

$\beta_h: parametro,\ \ para\ h = 1 \dots p\ dimensiones \qquad x_h: observasión,\ \ para\ h = 1 \dots p\ dimensiones$

$y_i: respuesta\ , \qquad para\ i = 1 \dots n\ observaciones$

Aunque es posible tener clases mal clasificadas, en la figura 4.3 se pueden apreciar gráficamente estos conceptos.

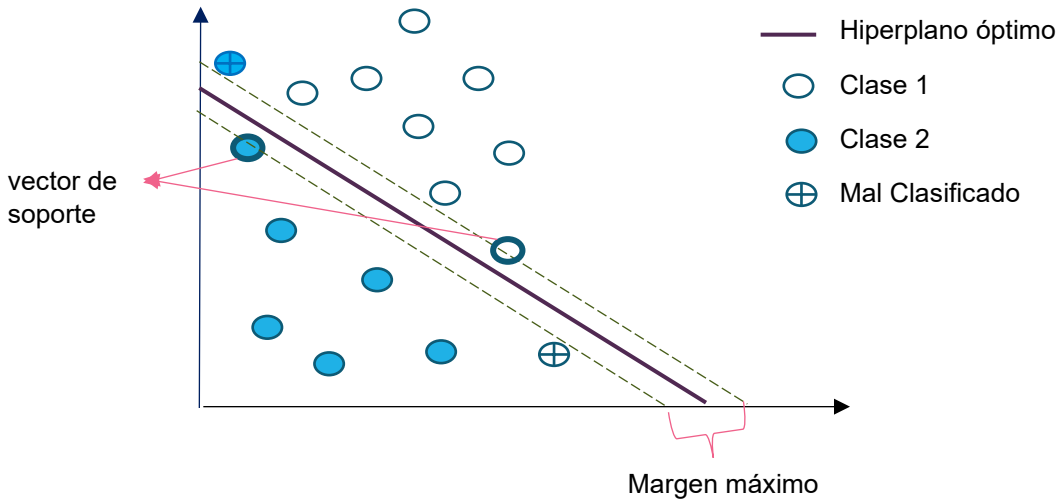


Figura 4.3 Ejemplo de dos clases separadas por un hiperplano

El hiperplano óptimo de la *figura 4.3.* que está definido por una línea recta pudiera ser representado por otras líneas y curvas, la forma de este hiperplano dependerá de la manera en la que sea óptimo separar los grupos. Las SVM permite generalizar el tipo de separación utilizando funciones núcleo (*kernel*) $K(\cdot)$. A su vez estas funciones núcleo pueden ser lineal, polinomial, radial, sigmoidal, entre otras.

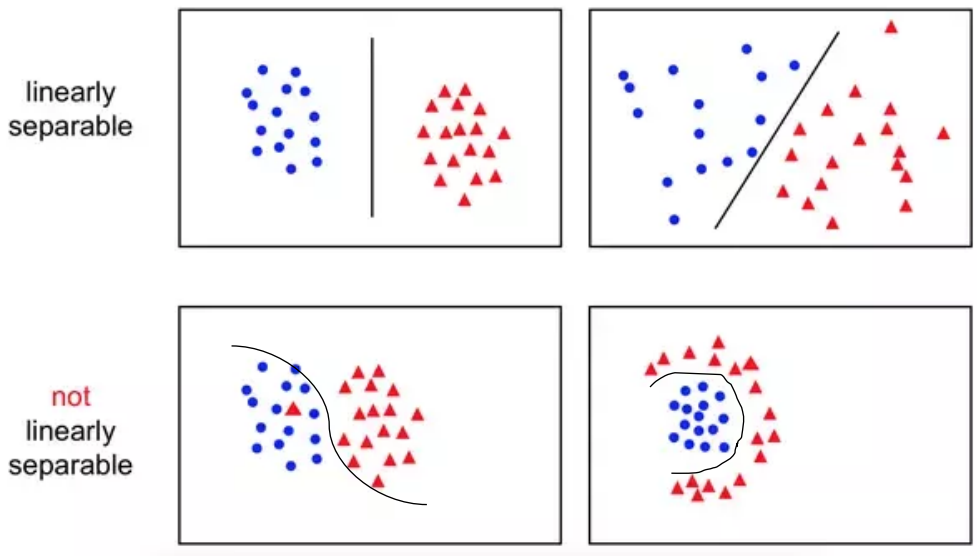


Figura 4.4 Ejemplo de funciones Kernel

https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

Un *kernel* $K(\cdot)$ en pocas palabras es una función que devuelve el resultado del producto al escalar entre dos vectores ejecutados en un nuevo espacio dimensional diferente al espacio

original en el que se encuentran los vectores. Si este producto escalar de la optimización es reemplazado por un *kernel*, los vectores de soporte (y el hiperplano) se obtienen directamente en el tamaño correspondiente al *kernel*. Con una pequeña modificación del problema original, a través de los *kernels*, el resultado se puede aplicar a cualquier dimensión.

La función kernel sigmoidal suele funcionar mejor para datos no balanceados ya que el resto tiende a categorizar los elementos solo en la clase mayoritaria, dicha función se define:

$$K(x, y) = \tan(\alpha x^t \cdot y + c), \qquad donde \ \ \alpha = \frac{1}{N} \ y \ c = constante$$

V. RESULTADOS

Para modelizar la probabilidad de que un asegurado cancele su póliza, se usa la base de datos analizada y detallada en apartado III del trabajo, es importante recalcar que las variables se transformaron a binarias para que los métodos de *machine learning* puedan entrenarse correctamente. Tanto para las NN como para las SVM se entrena un modelo con la base de datos de entrenamiento, dicho modelo se utiliza en la base de dato de prueba para examinar su rendimiento.

El rendimiento del modelo se evalúa comparando la información predicha con la observada a través de la matriz de confusión y sus diferentes medidas.

Como se desea comprobar la importancia de la siniestralidad a la hora de cancelar la póliza, se probaron los modelos óptimos con y sin la variable *pclaims* para todos los métodos.

5.1 APLICACIÓN MÉTODO NN-NEURAL NETWORK

Primero se entrenan las NN, tanto para las pólizas de auto como de las de hogar se toma como variable respuesta el estado de la póliza, para las variables explicativas se prueba agregar al modelo una a una hasta conseguir la combinación con el menor error, es decir, con la menor diferencia entre los datos de salida observados y los predichos.

En este trabajo se usa la función de *R* `nnet()`, la cual posee múltiples parámetros para su entrenamiento, sin embargo de la utiliza los siguientes parámetros:

- `formula:` Fórmula objetivo de forma: $y \sim x_1 \dots x_j, \quad para\ j = 1 \dots k\ categorías$
- `data:` `data.frame` o matriz de datos de origen.
- `size:` Número de nodos o neuronas intermedios.
- `maxit:` Número máximo de iteraciones, si pasado este número de búsquedas no se encuentra el peso óptimo la red no converge (se pudo ajustar).

5.1.1 Pólizas de auto

El modelo de NN con mejores resultados para predecir si el asegurado cancelará su póliza de auto o no, fue resultado de un estudio discriminatorio de variables explicativas, este modelo se probó con y sin la siniestralidad, de resto, las variables que se usaron sin la siniestralidad fueron:

- Edad del cliente
- Prima previa a la ultima
- Ultima prima
- Diferencia prima actual y previa
- Diferencia prima actual y primera
- Antigüedad del cliente
- Suplementos de la póliza
- Método de pago
- Tipo de vehículo
- Garantía contratada
- Potencia del vehículo
- Bono del cliente
- Número de asientos
- Valor del vehículo
- Pólizas adicionales

En total 16 de las 22 variables de la base de datos de las pólizas de auto fueron las utilizadas, resulta interesante que el sexo del asegurado no aportaba información relevante a la red, tampoco lo hizo la antigüedad del vehículo, el tipo de combustible, la antigüedad del permiso de conducir, ni si se tiene registrado un segundo conductor.

Modelo sin la siniestralidad:

```
NnModel20_a = nnet(policy_status_at_t~ Age_client+
previous_to_last_premium_paid+ last_premium_paid+ dif_current_previous+
dif_current_first+  Client_Seniority+ Policy_numSupplements
+Policy_PaymentMethod_A+car_type_M_Tour+contracted_guarantee_M_THIRD+
Car_power_M+ car_bonus_M +Car_number_of_seats_M
+value_of_car_9k+value_of_car_M+totalpol
, data=datos_entrena_a,size=17,maxit=1000,na.action = "na.omit")

## # weights:  307
## initial  value 169234.012790
## iter  10 value 165536.879753
…
## iter 360 value 112730.293377
## final  value 112728.955789
## converged
[1] 112729
```

Modelo con la siniestralidad:

```
NnModel20_a = nnet(policy_status_at_t ~ Age_client+
previous_to_last_premium_paid+ last_premium_paid+ dif_current_previous+
dif_current_first+  Client_Seniority+
Policy_numSupplements+Policy_PaymentMethod_A+car_type_M_Tour+contracted
_guarantee_M_THIRD+ Car_power_M+ car_bonus_M +Car_number_of_seats_M
+value_of_car_9k+value_of_car_M + pclaims, data=datos_entrena_a,
size=10, maxit=1000, na.action = "na.omit")

## # weights:  324
## initial  value 231172.333275
## iter  10 value 164313.180488
…
## iter 250 value 115969.063421final  value 115969.033898
## converged
[1] 115969
```

La NN asignó 307 pesos y convergió en la iteración 360 con un error 112729 en el modelo sin *pclaims*, y 324 pesos en el modelo con *pclaims* a en la interacción 250 y con un error de 115969.

A la NN con mejores resultados se le asignó 17 nodos igual, cantidad igual al número de variables explicativas, es usual utilizar este criterio para obtener el menor error posible, gráficamente se obtuvo:

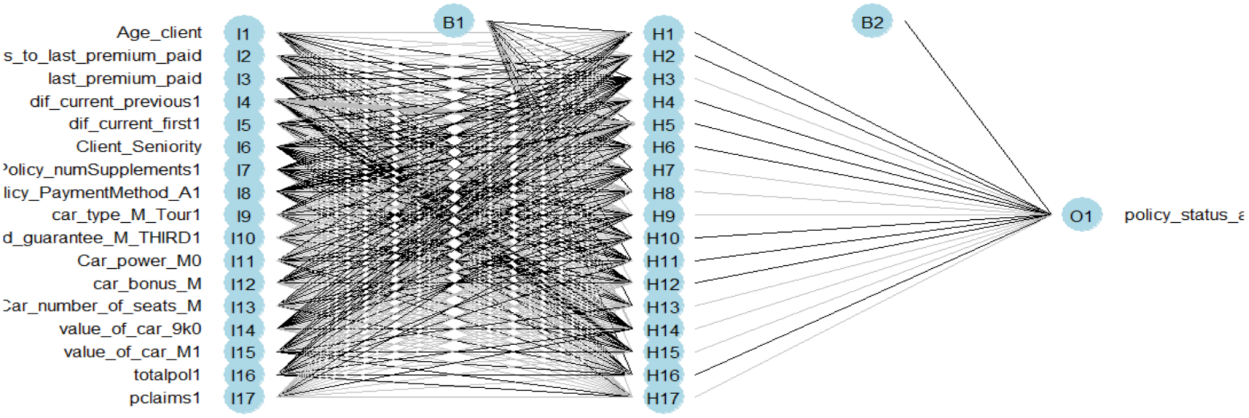


Figura 5.1 NN pólizas de auto con pclaims.

Una vez se obtuvo el modelo, se realizó la predicción en los datos de prueba teniendo en cuenta la desproporción del estado de la póliza de la base de datos de entrenamiento, dicha proporción se encuentra la tabla del anexo 9.1.1 y es del 80.4% para las pólizas que se encontraban vigentes.

Se evaluó el rendimiento a través de la matriz de confusión, los resultados obtenidos fueron:

Matriz de confusión sin pclaims

```
Confusion Matrix and Statistics

          obs_testa
pred_testa     0     1
         0 25395 27689
         1  3485 89816

               Accuracy : 0.787
                 95% CI : (0.7849, 0.7891)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4891

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8793
            Specificity : 0.7644
         Pos Pred Value : 0.4784
         Neg Pred Value : 0.9626
             Prevalence : 0.1973
         Detection Rate : 0.1735
   Detection Prevalence : 0.3626
      Balanced Accuracy : 0.8218

       'Positive' Class : 0
```

Tabla 5.1: Matriz de confusión auto sin pclaims

Matriz de confusión con pclaims

```
Confusion Matrix and Statistics

           obs_testap
pred_testap     0     1
          0 25113 26802
          1  3767 90703

               Accuracy : 0.7912
                 95% CI : (0.7891, 0.7933)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4931

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8696
            Specificity : 0.7719
         Pos Pred Value : 0.4837
         Neg Pred Value : 0.9601
             Prevalence : 0.1973
         Detection Rate : 0.1716
   Detection Prevalence : 0.3546
      Balanced Accuracy : 0.8207

       'Positive' Class : 0
```

Tabla 5.2: Matriz de confusión auto con pclaims

El modelo desarrollado predice aproximadamente el 78.7% de los datos sin la siniestralidad y 79.1% con ella, lo que significa si nos basamos solo en esta medida que la siniestralidad está ligeramente asociada a cancelar la póliza. Tanto la sensibilidad como la especificidad poseen valores altos, lo que significa que la predicción para las pólizas que se cancelan y las que no, es muy buena. Un dato que resalta el "*pos pred value*", que es la exactitud, se interpreta como que tan probable es que una póliza cancelada sea realmente cancelada, que con ambos modelos esa probabilidad es de 48%.

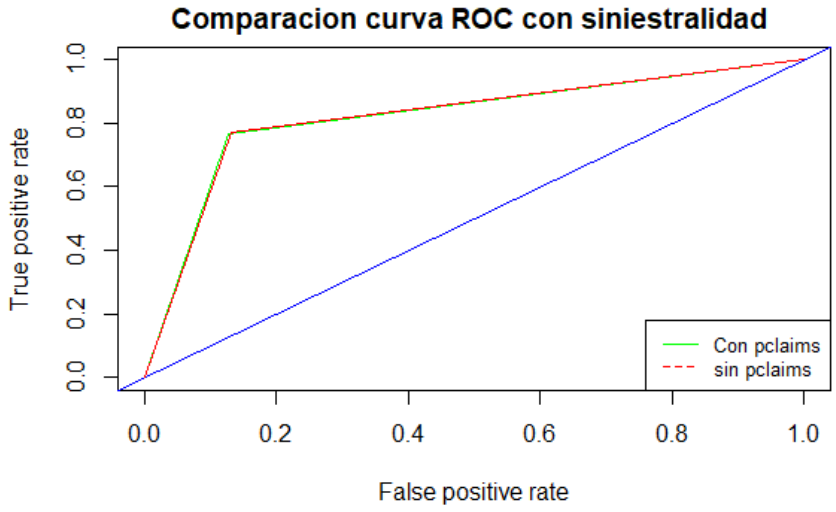


Figura 5.3 Curva ROC pólizas auto con y sin la siniestralidad.

La figura 5.3 muestra la curva ROC para los modelos con y sin la siniestralidad, ambas se posicionan por encima de la línea azul de aleatoriedad lo que gráficamente indica una buena predicción. También se calculó el AUC y resultó ser del 0.82 para ambas, insinuando que ambos modelos son buenos, aun así, se aprecia como la línea del modelo con la siniestralidad crece un poco más hacia la izquierda.

5.1.2 Pólizas de hogar

En el caso de la póliza del hogar resultó mejor utilizar las 13 variables como explicativas para predecir la probabilidad de que el asegurado cancele su póliza con 6 nodos intermedios, estas 13 variables sin la siniestralidad son:

- Sexo del cliente
- Edad del cliente
- Antigüedad del cliente
- Método de pago de la póliza
- Pago por la penúltima prima
- Pago por la ultima prima
- Diferencia entre la prima actual y la previa
- Diferencia entre la prima actual y la primera
- Pólizas adicionales
- Contenido de capital asegurado
- Continente capital asegurado
- Tipo de hogar

La asignación tanto de las variables como del número de nodo fue discriminante, resultó que el conjunto de todas las variables era necesario y que más de 6 nodos intermedios no mejoraban la predicción del modelo, los modelos resultantes fueron:

Modelo sin la siniestralidad:

```
NnModel1_h2 = nnet(policy_status_at_t~policy_status_at_t~
sex_customer + Age_client + Client_Seniority + Policy_PaymentMethod_A
+ dif_current_previous + dif_current_first + Insuredcapital_content_H
+ Insuredcapital_continent_H + totalpol + Policy_PaymentMethod_A +
HomeType_H_PI + pclaims , data=datos_entrena, size=6, maxit=10000,
na.action = "na.omit")

## # weights:  91
## initial  value 74993.740422
## iter  10 value 42230.798324
…
## iter 240 value 30424.177443
## final  value 30423.637487
## converged
```

Modelo con la siniestralidad:

```
NnModel20_h = nnet(policy_status_at_t~policy_status_at_t~ sex_customer +
Age_client + Client_Seniority + Policy_PaymentMethod_A +
dif_current_previous + dif_current_first + Insuredcapital_content_H +
Insuredcapital_continent_H + totalpol + Policy_PaymentMethod_A +
HomeType_H_PI + pclaims, data=datos_entrena, size=6, maxit=10000,
na.action = "na.omit")
# weights:  73
## initial  value 85408.654764
## iter  10 value 42233.981974
…
## iter  50 value 42228.686712
## final  value 42228.084103
## converged
```

La salida del modelo con *pclaims* informa que se asignaron 91 pesos a las variables, convergió en la iteración 240 y su error fue de 30423.64. En cambio, la salida sin *pclaims* asigna 73 pesos y converge en la iteracion 50 con un error de 42226.08.

Graficamente en la red los pesos mas significativos para la prediccion, se representan con lineas mas gruesas y oscura, en el caso del modelo con pclaims se tiene:

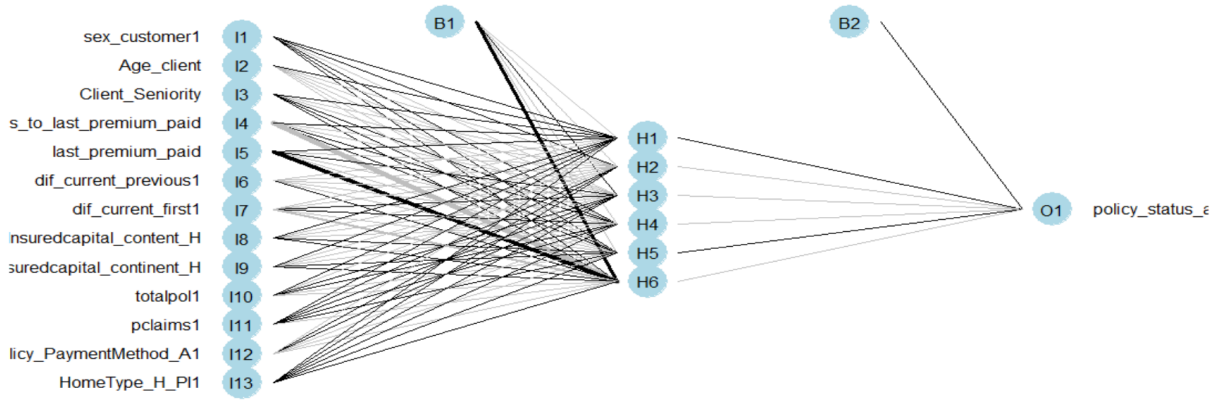


Figura 5.4 NN para las pólizas de hogar con pclaims

Su rendimiento se midió a través de los resultados de las siguientes matrices de confusión, teniendo en cuenta el desbalance de las clases del estado de la póliza, esta proporción se encuentra en la tabla anexa 9.1.2 y es del 89% para las pólizas no canceladas.

Tabla: Matriz de confusión sin pclaims

```
Confusion Matrix and Statistics

          obs_test_h
pred_test_h     0     1
          0  5186 41714
          1   494  4315

               Accuracy : 0.1837
                 95% CI : (0.1804, 0.1871)
    No Information Rate : 0.8902
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0016

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.91303
            Specificity : 0.09375
         Pos Pred Value : 0.11058
         Neg Pred Value : 0.89728
             Prevalence : 0.10985
         Detection Rate : 0.10029
   Detection Prevalence : 0.90700
      Balanced Accuracy : 0.50339

       'Positive' Class : 0
```

Tabla 5.4: Matriz de confusión hogar sin pclaims

Tabla: Matriz de confusión con pclaims

```
Confusion Matrix and Statistics

           obs_test_h2
pred_test_h2     0     1
           0  5619 16303
           1    61 29726

               Accuracy : 0.6835
                 95% CI : (0.6795, 0.6875)
    No Information Rate : 0.8902
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2818

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9893
            Specificity : 0.6458
         Pos Pred Value : 0.2563
         Neg Pred Value : 0.9980
             Prevalence : 0.1098
         Detection Rate : 0.1087
   Detection Prevalence : 0.4239
      Balanced Accuracy : 0.8175

       'Positive' Class : 0
```

Tabla 5.5: Matriz de confusión hogar con pclaims

El resultado de la predicción sin la siniestralidad resulta ser malo a primera impresión con un 18% de precisión, pero, tiene una sensibilidad del 91%, lo que significa que, si se desea solo evaluar la predicción de la clase positiva, que en este caso es que el asegurado cancele su póliza, se obtendría solo un 9% de cancelados observados que se predicen como no canceladas. Al agregar la siniestralidad el modelo mejora a un 68% de precisión, por lo que el modelo se puede considera como bueno con una alta sensibilidad.

Por último, la curva ROC de ambos modelos resultante es:

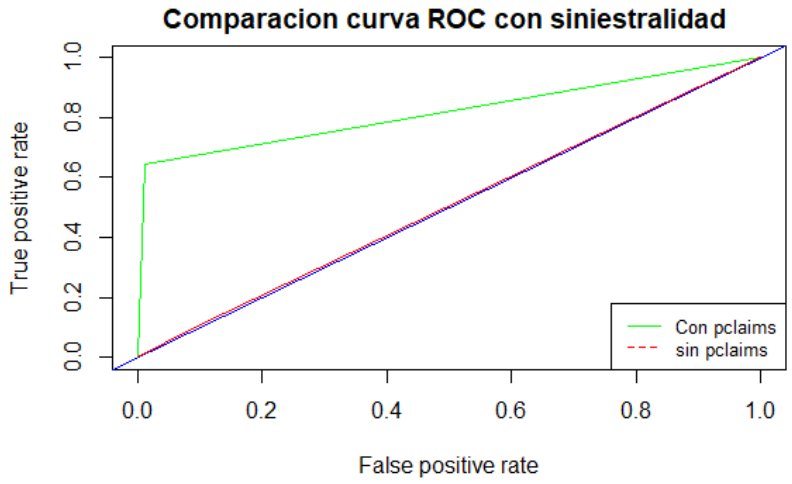


Figura 5.5 Curva ROC póliza hogar

Gráficamente resulta evidente que la curva del modelo sin la siniestralidad se posiciona sobre la recta de aleatoriedad azul, con un AUC igual al 0.5, no predice más que de lo que un modelo aleatorio de probabilidad 0.5 para cada clase, predicen prácticamente igual, lo que significa que el modelo es malo. En cambio, con la siniestralidad esta área de debajo de la curva es de 0.82, el modelo resulta bueno.

5.2 MÉTODO SVM SUPPORT VECTOR MACHINE

Ahora se entrenan las SVM con la misma fórmula objetivos de las NN. La función de *R* que se utiliza se llama `svm()`, como define (Bolance, 2020) «la función no posee gran interés a nivel de interpretación, simplemente informa del tipo de modelo, que en este caso es el `C-classification`, de algunos parámetros utilizados en el proceso y del número de iteraciones.»

Como se mencionó en la explicación de las SVM se usa `kernel="sigmoid"`, porque según estudios anteriores tiende a no clasificar solo en la clase más frecuente.

Lo importante de la función es la predicción que se obtiene a partir ese modelo. El resultado del modelo es la categoría predicha a la que pertenecen los datos, es decir, que los resultados son discretos, 0 si predice que cancela la póliza, 1 si no la cancela, por tanto, gráficamente no da más información la “curva” ROC, ya que sería un punto en el plano.

5.2.1 Pólizas de auto

En las NN se definieron las variables explicativas utilizadas para modelar la probabilidad de cancelar la póliza de auto, en las SVM se utilizaron las misma, la salida los modelos resultantes fueron:

En el modelo sin la siniestralidad necesitó de 108516 vectores para clasificar si el asegurado cancela su póliza o no, y la salida resultan luce:

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
      coef.0:  0

Number of Support Vectors:  108516
```

El modelo con la siniestralidad necesito de 108440 vectores para clasificar el estado de la póliza el resultado de la función es:

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
     coef.0:  0

Number of Support Vectors:  108440
```

La matriz de confusión de las pólizas predichas versus las observadas se utiliza igualmente para evaluar la capacidad predictiva de las SVM, a continuación, se muestran con y sin la siniestralidad:

Matriz de confusión sin pclaims

```
Confusion Matrix and Statistics

predX1_TestSVM_a     0     1
               0  5512 23349
               1 23368 94156

               Accuracy : 0.6809
                 95% CI : (0.6785, 0.6832)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.0078

 Mcnemar's Test P-Value : 0.9336

            Sensitivity : 0.19086
            Specificity : 0.80129
         Pos Pred Value : 0.19098
         Neg Pred Value : 0.80116
             Prevalence : 0.19729
         Detection Rate : 0.03765
   Detection Prevalence : 0.19716
      Balanced Accuracy : 0.49608

       'Positive' Class : 0
```

Tabla 5.6: Matriz de confusión auto sin pclaims

Matriz de confusión con pclaims

```
Confusion Matrix and Statistics

                obs_test_SVM_a2
pred_test_SVM_a2     0     1
               0  5531 23303
               1 23349 94202

               Accuracy : 0.6813
                 95% CI : (0.6789, 0.6837)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : 1.000

                  Kappa : -0.0068

 Mcnemar's Test P-Value : 0.835

            Sensitivity : 0.19152
            Specificity : 0.80169
         Pos Pred Value : 0.19182
         Neg Pred Value : 0.80137
             Prevalence : 0.19729
         Detection Rate : 0.03778
   Detection Prevalence : 0.19697
      Balanced Accuracy : 0.49660

       'Positive' Class : 0
```

Tabla 5.7: Matriz de confusión auto con pclaims

La matriz de confusión en el modelo de las SVM sin la siniestralidad predice aproximadamente un 0.2% menos que el modelo con la siniestralidad. Las predicciones tienen un 68% de precisión, al inicio no parece un valor tan malo, pero, al analizar la sensibilidad y la exactitud resulta ser muy baja, del 19% para ambas lo que decir que el modelo no predice correctamente a los asegurados que cancelan su póliza, en cambio, para los que no la cancelan la especificidad alcanza el 80%, este caso no resulta tan interesante para retener a los asegurados, sin embargo, pudiera servir en otro planteamiento.

En las siguientes gráficas, los puntos que están representados por una “x” son los vectores de soporte, son los puntos que afectan directamente la línea de clasificación y los puntos marcados con una "o" son los otros puntos, que no afectan el cálculo de la línea.

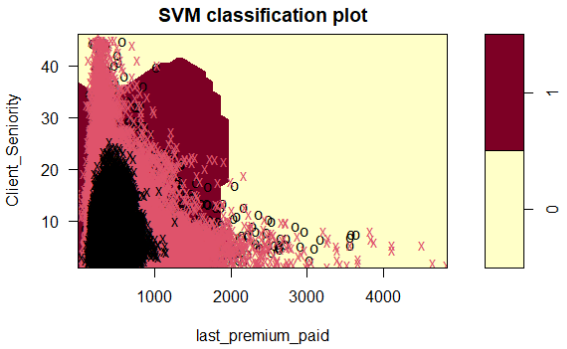


Figura 5.6 SVM A antigüedad y ultima prima sin pclaims

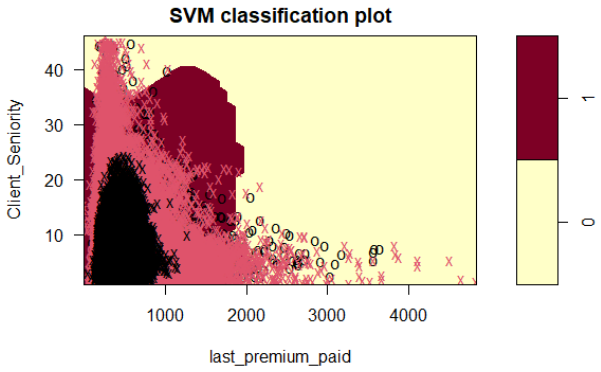


Figura 5.7 SVM A antigüedad y ultima prima con pclaims

Al graficar la SVM para las variables antigüedad del cliente con la última prima pagada para ambos modelos resultan muy similares, la línea de separación es curva. El color crema es la parte del plano que representa a las pólizas canceladas y el color rojo a las no canceladas, cuya área parece indicar que hay una concentración de predicciones de pólizas no canceladas, mientras que las canceladas se sitúan en los espacios extremos superiores del plano.

5.2.2 Pólizas de hogar

Para predecir si se cancelara la póliza de hogar o no, se usó igual que en las NN, todas las variables, primero sin la siniestralidad y luego con ella. El resultado de la función en R fue:

El modelo sin la siniestralidad:

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
      coef.0:  0

Number of Support Vectors:  108516
```

La máquina de soporte necesitó de 108516 vectores para clasificar si el asegurado cancelará su póliza o no.

El modelo de las SVM con la siniestralidad:

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
     coef.0:  0

Number of Support Vectors:  108440
```

La máquina de soporte necesitó de 108440 vectores para clasificar el estado de la póliza.

Para evaluar la capacidad predictiva del modelo de SVM, se utilizan las siguientes tablas de confusión:

Tabla: Matriz de confusión sin pclaims

```
Confusion Matrix and Statistics

predX1_TestSVM_h     0     1
               0   694  4895
               1  4986 41134

               Accuracy : 0.8089
                 95% CI : (0.8055, 0.8123)
    No Information Rate : 0.8902
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0159

 Mcnemar's Test P-Value : 0.3653

            Sensitivity : 0.12218
            Specificity : 0.89365
         Pos Pred Value : 0.12417
         Neg Pred Value : 0.89189
             Prevalence : 0.10985
         Detection Rate : 0.01342
   Detection Prevalence : 0.10809
      Balanced Accuracy : 0.50792

       'Positive' Class : 0
```

Tabla 5.8: Matriz de confusión hogar sin pclaims

Tabla: Matriz de confusión con pclaims

```
Confusion Matrix and Statistics

                 obs_test_SVM_h2
pred_test_SVM_h2     0     1
               0   713  4892
               1  4967 41137

               Accuracy : 0.8093
                 95% CI : (0.8059, 0.8127)
    No Information Rate : 0.8902
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0194

 Mcnemar's Test P-Value : 0.4561

            Sensitivity : 0.12553
            Specificity : 0.89372
         Pos Pred Value : 0.12721
         Neg Pred Value : 0.89227
             Prevalence : 0.10985
         Detection Rate : 0.01379
   Detection Prevalence : 0.10840
      Balanced Accuracy : 0.50962

       'Positive' Class : 0
```

Tabla 5.9: Matriz de confusión hogar con pclaims

La precisión del modelo más simple es del 80.89%, un 0.04% menos que el SVM con siniestralidad. A primera impresión un buen resultado, pero cuando se profundiza en las otras medidas como la sensibilidad, esta es de solo un 12.5% en el modelo más complejo, lo que significa que las pólizas no canceladas no se clasifican correctamente, el porcentaje de pólizas canceladas bien clasificadas se encuentra en la exactitud que también es del 12%, por lo cual se pierde el 88% de información para conocer realmente la probabilidad de que un asegurado cancele su póliza.

Si se grafica la antigüedad del cliente con el continente de capital asegurado, se ve como el hiperplano separa en un pequeño intervalo del capital asegurado la información de las pólizas

vigentes, dejando como pólizas canceladas los valores más extremos del continente asegurado, mientras que la antigüedad no aparenta tener variación según el estado de la póliza.

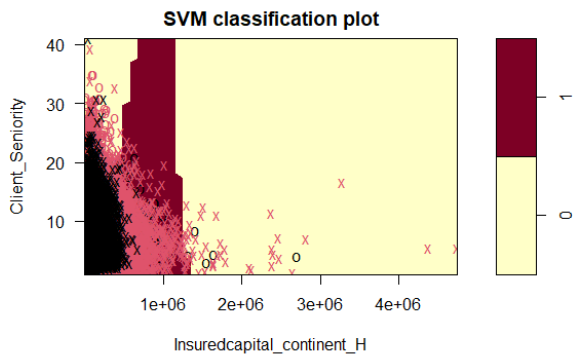


Figura 5.8 SVM antigüedad y capital asegurado sin pclaims

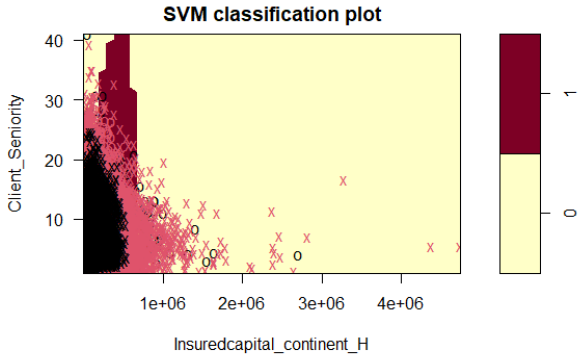


Figura 5.9 SVM antigüedad y capital asegurado con pclaims

VI. DISCUSIÓN

6.1 COMPARACIÓN DE MÉTODOS

En los métodos de predicción tradicionales para las ciencias actuariales, como los modelos lineales o los GML se suele usar AIC (criterio de información Akaike) o BIC (criterio de información bayesiano) para comparar los modelos, pero, para los métodos de *machine learning* estos criterios no son aptos, ya que, ambos se basan en la verosimilitud y no en el error de las NN o el margen máximo de las SVM. Por eso, la comparación entre las NN y las SVM se basan en la matriz de confusión desde el punto de vista de la autora del trabajo.

Para conocer la proporción de pólizas que serán canceladas en una matriz de datos siendo la clase “positiva”, en este caso, que el asegurado cancele su póliza, las medidas que la autora considera más importantes, más allá de la precisión, es la sensibilidad por el hecho de que es la proporción de pólizas canceladas predichas con respecto a las observadas. A la hora de tomar alguna acción ejecutiva preventiva con la proporción de pólizas que se predicen como canceladas, es mejor tener en cuenta todas las predichas, aunque estas incluyan falsos positivos, es decir, pólizas no canceladas que se predicen como canceladas, por el hecho de ser una medida preventiva que favorecerá al cliente no tendrían porqué verse afectados negativamente estos falsos positivos. En caso de que se quiera evitar estos falsos negativos, la siguiente medida más importante diría, que es la exactitud por ser la proporción de pólizas canceladas observadas con respecto a las predichas.

6.1.1 Pólizas de auto

Por tanto, a lo explicado en el párrafo anterior, al comparar los métodos de *machine learning* con y sin la siniestralidad, el modelo que resulta mejor es el de NN, aunque no hay mayor diferencia entre las NN con y sin la siniestralidad, por el hecho de que declarar siniestros es un factor relevante en otros estudios seleccionaría el modelo con la siniestralidad.

Otro aspecto que resaltar es el coste y tiempo computacional, las NN fueron mucho más rápidas de entrenar con respecto a las SVM, el hecho de asignar 324 pesos resultó ser más rápido que crear 108440 vectores.

Tabla de comparación de los modelos de las pólizas de auto

Medida	Neural Network con pclaims	Neural Network sin pclaims	Support Vector Machine con pclaims	Support Vector Machine sin pclaims
Precisión	0.7912	0.7870	0.6809	0.6813
Sensibilidad	0.8696	0.8793	0.1909	0.1915
Especificidad	0.7719	0.7644	0.8013	0.8017
Exactitud	0.4837	0.4784	0.1910	0.1918

Tabla 6.1 Comparación de modelos en pólizas de auto

En este caso el modelo seleccionado es la NN con la siniestralidad, con un 79% de precisión, 87% de sensibilidad y 48% de exactitud. Lo que significa que se predijo el 87% de las pólizas canceladas de las cuales, el 48% son realmente canceladas.

6.1.2 Pólizas de hogar

En el caso de las pólizas de hogar al haber menor cantidad de observaciones la programación de ambos modelos tomo un tiempo similar.

Sin embargo, el resultado del mejor modelo coincide con el de las pólizas de auto, aunque aquí la siniestralidad jugó un papel de gran importancia mejorando la precisión en un 40% el modelo sin ella.

Tabla de comparación de los modelos de las pólizas de hogar

Medida	Neural Network con pclaims	Neural Network sin pclaims	Support Vector Machine con pclaims	Support Vector Machine sin pclaims
Precisión	0.6835	0.1837	0.8089	0.8093

Sensibilidad	0.9893	0.9130	0.1222	0.1255
Especificidad	0.6458	0.0938	0.8937	0.8937
Exactitud	0.2563	0.1106	0.1242	0.1272

Tabla 6.2 Comparación de modelos en pólizas de hogar

Lo que significa que le modelo seleccionado es el la NN con la siniestralidad, con una precisión del 68%, una sensibilidad alta que alcanza el 99% pero, con la exactitud del 25%. Lo que significa que se predijo el 99% de las pólizas canceladas de las cuales, el 26% son realmente canceladas.

6.2 COMPARACIÓN TIPOS DE PÓLIZAS

Con respecto a las variables en común entre las pólizas de hogar y de auto, que son las características de la póliza y del cliente, resultaron importantes al momento de la predicción, aún así, resulta interesante como en el caso de las pólizas de auto, el sexo del cliente no tuvo ninguna influencia ni mejora para el modelo, en cambio, en las pólizas de hogar si fue relevante. Otra variable que resaltar es la importancia que tuvo la siniestralidad en las pólizas del hogar mejorando significativamente el modelo, mientras que en las pólizas de auto no tuvo mayor importancia.

Variable	Descripción	Auto	Hogar
totalpol	Pólizas adicionales	X	X
sex_customer	Sexo del cliente		X
Age_client	Edad del cliente	X	X
Client_Seniority	Antigüedad del cliente	X	X
Policy_PaymentMethod	Método de pago de la póliza	X	X
previous_to_last_premium_paid	Pago por la penúltima prima	X	X
last_premium_paid	Pago por la ultima prima	X	X
dif_current_previous	Diferencia entre la prima actual y la previa	X	X
dif_current_first	Diferencia entre la prima actual y la primera	X	X
Pclaims	Siniestros declarados	X	X

Tabla 6.3 Comparación de variables en pólizas de auto vs hogar

Aunque el modelo para ambas fueron las NN, en las pólizas de auto estas tuvieron un mejor rendimiento que en las de hogar, puede ser por múltiples factores como la naturaleza de los datos, el número de variables explicativas, el número de observaciones, lo valores extremos del capital asegurado o la combinación de todas, pero la exactitud del modelo para las pólizas de auto duplico a la de las pólizas del hogar.

VII. CONCLUSIONES

Este trabajo compara el desempeño de diferentes técnicas *machine learning* de clasificación y predicción con respecto a la retención del asegurado en las compañías de seguros para pólizas de no vida, auto y hogar. Tras un estudio de la matriz de datos y de los métodos se llega a la conclusión que, a pesar de que tanto las NN como las SVM tienen una precisión sobre el 70%, aun así las NN se pueden adaptarse mejor a la problemática de la desproporción del estado de las pólizas de los datos observados con una sensibilidad de más del 80%, en cambio las SVM tienden a favorecer a las pólizas no canceladas por ser la clase mayoritaria y clasificar como canceladas solo a los valores extremos de los datos, lo que da como resultado una sensibilidad mala, debajo del 20% en ambos tipos de pólizas.

Frente a la evidencia recaudada de las variables predictoras, hacer una limpieza y análisis previo al entrenamiento fue indispensable para conseguir una buena modelización. Tanto las características contractuales de la póliza y su histórico como las del cliente, tuvieron un papel indispensable a la hora de explicar el estado de la póliza, exceptuando el sexo para las pólizas de auto. Con respecto al objeto asegurado su valor y tipología, en ambos tipos de póliza contribuyen a explicar el estado de estas, sin embargo, otras características propias del auto como el tipo de combustible, tener un segundo conductor, la antigüedad del auto y del permiso de conducir no contribuyeron.

En especial, la siniestralidad tuvo una gran importancia a la hora de clasificar el estado de las pólizas, específicamente para las pólizas de hogar, mejorando el desempeño del modelo un 40% más de precisión que el modelo sin ella.

Los resultados obtenidos, aunque no sean exactamente extrapolables, se pueden utilizar en la cartera origen de pólizas y probar en alguna diferente para conocer la proporción de pólizas con más posibilidades de ser cancelas, con el fin de empelar acciones de retención para evitar la pérdida de clientes.

VIII. BIBLIOGRAFIA

[1]. BISHOP, Christopher M. Pattern recognition. *Machine learning*, 2006, vol. 128, no 9.

[2]. CHALAPATHY, Raghavendra; CHAWLA, Sanjay. *Deep learning for anomaly detection: A survey*. *arXiv preprint arXiv:1901.03407*, 2019. (https://doi.org/10.1016/j.jnca.2016.04.007)

[3]. DELICADO, Pedro; PEÑA, *Daniel. Understanding complex predictive models with Ghost Variables*. *arXiv preprint arXiv:1912.06407*, 2019.

[4]. FERRER, Virginia, et. al. (eds.). *El trabajo fin de Grado. Guía para estudiantes, docentes y agentes colaboradores.* Aravaca: McGraw-Hill, 201

[5]. GOODFELLOW, Ian, et al. *Deep learning*. Cambridge: MIT press, 2016.

[6]. MURPHY, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[7]. PORRAS, Salvador Torra. *La Siniestralidad en seguros de consumo anual de las entidades de previsión social: Perspectiva probabilística y econométrica. Propuesta de un modelo econométrico neuronal para Cataluña*. 2004. Tesis Doctoral. Universitat de Barcelona.

[8]. RIPLEY, Brian; VENABLES, William; RIPLEY, Maintainer Brian. Package 'nnet'. *R package version*, 2016, vol. 7, p. 3-12.

[9]. RODRIGO, J. Amat. Máquinas de Vector Soporte (Support Vector Machines, SVMs). 2017. (https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines)

IX. ANEXOS

9.1 TABLAS COMPARATIVAS BASE DE DATOS ENTRENAMIENTO Y TEST

9.1.1 Auto

	Entrenamiento (N=341225)	Prueba (N=146385)	Total (N=487610)
sex_customer			
0	241838 (70.9%)	104024 (71.1%)	345862 (70.9%)
1	99387 (29.1%)	42361 (28.9%)	141748 (29.1%)
Age_client			
Mean (SD)	52.3 (13.5)	52.2 (13.5)	52.3 (13.5)
Median [Min, Max]	52.0 [18.0, 99.0]	52.0 [18.0, 99.0]	52.0 [18.0, 99.0]
Client_Seniority			
Mean (SD)	5.43 (4.40)	5.42 (4.40)	5.42 (4.40)
Median [Min, Max]	4.12 [1.00, 46.2]	4.10 [1.00, 51.8]	4.11 [1.00, 51.8]
Policy_numSupplements			
0	75122 (22.0%)	32570 (22.2%)	107692 (22.1%)
1	266103 (78.0%)	113815 (77.8%)	379918 (77.9%)
previous_to_last_premium_paid			
Mean (SD)	386 (199)	385 (200)	386 (199)
Median [Min, Max]	329 [2.80, 4830]	328 [17.3, 6540]	329 [2.80, 6540]
last_premium_paid			
Mean (SD)	392 (204)	392 (204)	392 (204)
Median [Min, Max]	333 [4.04, 4830]	333 [17.3, 6540]	333 [4.04, 6540]
dif_current_previous			
0	87553 (25.7%)	37580 (25.7%)	125133 (25.7%)
1	253672 (74.3%)	108805 (74.3%)	362477 (74.3%)
dif_current_first			
0	146492 (42.9%)	63293 (43.2%)	209785 (43.0%)
1	194733 (57.1%)	83092 (56.8%)	277825 (57.0%)
age_of_car_M			
Mean (SD)	10.6 (5.63)	10.6 (5.65)	10.6 (5.64)
Median [Min, Max]	10.0 [0, 68.0]	10.0 [0, 88.0]	10.0 [0, 88.0]
Car_Years1stDriverLicense_M			
Mean (SD)	28.9 (12.5)	28.8 (12.5)	28.9 (12.5)
Median [Min, Max]	28.0 [0, 79.0]	28.0 [0, 74.0]	28.0 [0, 79.0]
Car_number_of_seats_M			

	Entrenamiento (N=341225)	Prueba (N=146385)	Total (N=487610)
0	49231 (14.4%)	21014 (14.4%)	70245 (14.4%)
1	291994 (85.6%)	125371 (85.6%)	417365 (85.6%)
Car_power_M			
1	254737 (74.7%)	109374 (74.7%)	364111 (74.7%)
0	86488 (25.3%)	37011 (25.3%)	123499 (25.3%)
car_bonus_M			
1	75122 (22.0%)	32570 (22.2%)	107692 (22.1%)
2	266103 (78.0%)	113815 (77.8%)	379918 (77.9%)
Car_2ndDriver_M			
0	298439 (87.5%)	128150 (87.5%)	426589 (87.5%)
1	42786 (12.5%)	18235 (12.5%)	61021 (12.5%)
value_of_car_M			
0	139164 (40.8%)	59775 (40.8%)	198939 (40.8%)
1	202061 (59.2%)	86610 (59.2%)	288671 (59.2%)
policy_status_at_t			
0	66849 (19.6%)	28880 (19.7%)	95729 (19.6%)
1	274376 (80.4%)	117505 (80.3%)	391881 (80.4%)
totalpol			
0	312977 (91.7%)	134131 (91.6%)	447108 (91.7%)
1	28248 (8.3%)	12254 (8.4%)	40502 (8.3%)
pclaims			
0	326149 (95.6%)	139926 (95.6%)	466075 (95.6%)
1	15076 (4.4%)	6459 (4.4%)	21535 (4.4%)
Policy_PaymentMethod_A			
0	82486 (24.2%)	35538 (24.3%)	118024 (24.2%)
1	258739 (75.8%)	110847 (75.7%)	369586 (75.8%)
contracted_guarantee_M_THIRD			
0	71017 (20.8%)	30177 (20.6%)	101194 (20.8%)
1	270208 (79.2%)	116208 (79.4%)	386416 (79.2%)
car_type_M_Tour			
0	75692 (22.2%)	32245 (22.0%)	107937 (22.1%)
1	265533 (77.8%)	114140 (78.0%)	379673 (77.9%)
Fuel_Type_M_D			
0	147155 (43.1%)	63088 (43.1%)	210243 (43.1%)
1	194070 (56.9%)	83297 (56.9%)	277367 (56.9%)
value_of_car_9k			
1	6185 (1.8%)	2688 (1.8%)	8873 (1.8%)

	Entrenamiento (N=341225)	Prueba (N=146385)	Total (N=487610)
0	335040 (98.2%)	143697 (98.2%)	478737 (98.2%)
muestra			
Mean (SD)	1.00 (0)	2.00 (0)	1.30 (0.458)
Median [Min, Max]	1.00 [1.00, 1.00]	2.00 [2.00, 2.00]	1.00 [1.00, 2.00]

9.1.2 Hogar

	Entrenamiento (N=120999)	Prueba (N=51709)	Total (N=172708)
sex_customer			
0	86212 (71.3%)	36840 (71.2%)	123052 (71.2%)
1	34787 (28.7%)	14869 (28.8%)	49656 (28.8%)
Age_client			
Mean (SD)	59.7 (13.6)	59.7 (13.6)	59.7 (13.6)
Median [Min, Max]	60.0 [18.0, 85.0]	60.0 [18.0, 85.0]	60.0 [18.0, 85.0]
Client_Seniority			
Mean (SD)	6.67 (4.37)	6.67 (4.38)	6.67 (4.37)
Median [Min, Max]	5.65 [1.00, 41.1]	5.63 [1.00, 40.4]	5.65 [1.00, 41.1]
previous_to_last_premium_paid			
Mean (SD)	253 (170)	252 (171)	252 (170)
Median [Min, Max]	207 [39.6, 4870]	207 [0, 5600]	207 [0, 5600]
last_premium_paid			
Mean (SD)	250 (170)	250 (172)	250 (171)
Median [Min, Max]	204 [38.6, 4980]	205 [37.6, 5600]	204 [37.6, 5600]
dif_current_previous			
0	17853 (14.8%)	7599 (14.7%)	25452 (14.7%)
1	103146 (85.2%)	44110 (85.3%)	147256 (85.3%)
dif_current_first			
0	6805 (5.6%)	2908 (5.6%)	9713 (5.6%)
1	114194 (94.4%)	48801 (94.4%)	162995 (94.4%)
Insuredcapital_content_H			
Mean (SD)	34800 (34700)	34800 (33600)	34800 (34400)
Median [Min, Max]	27900 [1.03, 1640000]	28000 [165, 1010000]	28000 [1.03, 1640000]
Insuredcapital_continent_H			
Mean (SD)	123000 (110000)	123000 (114000)	123000 (111000)
Median [Min, Max]	96200 [180, 4740000]	96300 [225, 4400000]	96200 [180, 4740000]
policy_status_at_t			

	Entrenamiento (N=120999)	Prueba (N=51709)	Total (N=172708)
0	13457 (11.1%)	5680 (11.0%)	19137 (11.1%)
1	107542 (88.9%)	46029 (89.0%)	153571 (88.9%)
totalpol			
0	99413 (82.2%)	42389 (82.0%)	141802 (82.1%)
1	21586 (17.8%)	9320 (18.0%)	30906 (17.9%)
Policy_PaymentMethod_A			
0	7986 (6.6%)	3535 (6.8%)	11521 (6.7%)
1	113013 (93.4%)	48174 (93.2%)	161187 (93.3%)
HomeType_H_PI			
0	43665 (36.1%)	18631 (36.0%)	62296 (36.1%)
1	77334 (63.9%)	33078 (64.0%)	110412 (63.9%)

9.2 CÓDIGO DE R

```
oridinal <- read_sas("final2013.sas7bdat", NULL)
final2013<-oridinal
#Variables respuestas
final2013$policy_status_at_t<-factor(revalue(final2013$policy_status_at_t, c("V"=1, "A"
=0)))

#PREPROSESING

#suma de variables
final2013$totalclaims<-(final2013$nclaims_bi_ins_C+final2013$nclaims_md_ins_C+final2013
$nclaims_home_ins_C)

final2013$totalpol <-(final2013$nunpol_accidents+final2013$nunpol_InsRetPlan+final2013$
nunpol_life_risk+final2013$nunpol_life_saving+final2013$nunpol_other)

#eliminacion filas no necesarias
final2013<- final2013[,-c(1:3,5:7,9:13,15:17,25,27,37,38,44:55)]

#outlier
outlier2<- which(abs(final2013$dif_current_first)>1000)
final2013<-final2013[-outlier2,]
outlier3<- which(abs(final2013$dif_current_previous)>1000)
final2013<-final2013[-outlier3,]

#Categorizacion variables binarias
final2013$sex_customer<- factor(revalue(final2013$sex_customer, c("Male"=0, "Female"=1)
))

final2013$Car_2ndDriver_M<- factor(revalue(final2013$Car_2ndDriver_M, c("NO"=0, "YES"=1
)))

final2013$dif_current_first<-ifelse((final2013$dif_current_first >= 0), 1, 0)
final2013$dif_current_first<-as.factor(final2013$dif_current_first)
```

```
final2013$dif_current_previous<-ifelse((final2013$dif_current_previous>= 0), 1, 0)
final2013$dif_current_previous<-as.factor(final2013$dif_current_previous)

#categorizacion variables categoricas:
final2013$Policy_PaymentMethod_A<-factor( revalue(final2013$Policy_PaymentMethod,c("A"=
1,"S"=0,"T"=0,"U"=0)))

final2013$contracted_guarantee_M_THIRD<-factor( revalue(final2013$contracted_guarantee_
M,c("THIRD PARTY"=1,"ALL RISKS WITHOUT FRA"=0,"ALL RISKS WITH FRA"=0)))
final2013$car_type_M_Tour<- factor(revalue(final2013$car_type_M,c(
"TOURISM"=1,"MINIVAN"=0, "ALL TERRAIN"=0)))

final2013$Fuel_Type_M_D<- factor(revalue(final2013$Fuel_Type_M,c(
"Gasoline"=0,"Diesel"=1,"Hybrid"=0 )))

final2013$Car_number_of_seats_M<- ifelse((final2013$Car_number_of_seats_M== "5"), 1, 0)

final2013$totalpol<- ifelse((final2013$totalpol== "0"), 0, 1)
final2013$totalpol<-as.factor(final2013$totalpol)

final2013$car_bonus_M<- ifelse((final2013$car_bonus_M== "0"), 0, 1)
final2013$car_bonus_M<-as.factor(final2013$car_bonus_M)

final2013$Policy_numSupplements<- ifelse((final2013$car_bonus_M== "0"), 0, 1)
final2013$Policy_numSupplements<-as.factor(final2013$Policy_numSupplements)

final2013$HomeType_H_PI<- factor(revalue(final2013$HomeType_H, c("AT"=0,"PB"=0,"PI"=1,"
RU"=0,"UA"=0,"UF"=0)))
#ELIMINACION VARIABLES BASE
final2013<- final2013[,-c(4,5,16,18,20,23,27)]

#arreglo variables NUMERICAS
final2013$value_of_car_M[final2013$value_of_car_M == "'0.0'"] <- "0k-1k"
final2013$value_of_car_M<- as.factor(final2013$value_of_car_M)
final2013$value_of_car_M <- ordered(final2013$value_of_car_M, levels = c( "",'0k-1k',
'1k-2k','2k-3k','3k-4k','4k-5k','5k-6k','6k-7k','7k-8k','8k-9k','9k-10k','10k-11k',
'11k-12k','12k-13k','13k-14k','14k-15k','15k-16k','16k-17k','17k-18k','18k-19k','19k-20
k',
'20k-21k','21k-22k','22k-23k','23k-24k','24k-25k','25k-26k','26k-27k','27k-28k','28k-29
k',
'29k-30k','30k-32k','32k-34k','34k-36k','36k-38k','38k-40k','40k-45k','45k-50k','50k-60
k',
'60k-70k','70k-80k','80k-90k','90k-100k','100k-high'))
final2013$value_of_car_M<- as.numeric(final2013$value_of_car_M)
set.seed(1234)
final2013$value_of_car_M<-discretize(final2013$value_of_car_M, "cluster", categories=3)
levels(final2013$value_of_car_M)
final2013$value_of_car_M<-factor(final2013$value_of_car_M)
final2013$value_of_car_9k<- factor(revalue(final2013$value_of_car_M,c(
"[1,8.68)"="1", "[8.68,22.9)"="0", "[22.9,44]"="0")))
final2013$value_of_car_M<- factor(revalue(final2013$value_of_car_M,c(
"[1,8.68)"="0", "[8.68,22.9)"="1", "[22.9,44]"="0")))

final2013$Car_power_M<-discretize(final2013$Car_power_M, "cluster", categories=2)
final2013$Car_power_M<- factor(final2013$Car_power_M)
final2013$Car_power_M<- factor(revalue(final2013$Car_power_M,c(
"[4,128)"="1", "[128,740]"="0")))

#outlier policy auto
outlier<- which(final2013$previous_to_last_premium_paid>174000 & final2013$policy_group
=='1')
final2013<-final2013[-outlier,]
```

```
summary(final2013)
auto<-final2013[final2013$policy_group=='1',-c(16:19,21,27)]
auto$car_bonus_M[auto$car_bonus_M == ""] <- "0"
auto$car_bonus_M<- as.numeric(auto$car_bonus_M)
auto$Fuel_Type_M_D<- factor(auto$Fuel_Type_M_D)
auto$car_type_M_Tour<- factor(auto$car_type_M_Tour)
auto$Car_2ndDriver_M<- factor(auto$Car_2ndDriver_M)
auto$contracted_guarantee_M_THIRD<- factor(auto$contracted_guarantee_M_THIRD)
nas8<-as.data.frame(which(is.na(auto),arr.ind=TRUE))
auto<- auto[-nas8$row,]
summary(auto)
hogar<-final2013[final2013$policy_group=='2',-c(4,9:15,18,19,21,23:26,28)]
hogar$HomeType_H_PI<- factor(hogar$HomeType_H_PI)

nas7<-as.data.frame(which(is.na(hogar),arr.ind=TRUE))
hogar<- hogar[-nas7$row,]
summary(hogar)
# Definimos una semilla para garantizar que cada vez que ejecutemos siempre vamos a obt
ener las mismas muestras
set.seed(1234)
#Generación de la muestras 70% vs 30% # Contamos el número de casos
n<-nrow(hogar)
# Definimos una variable que se llama muestra que nos identificaran los casos que se ti
enen que utilizar como entrenamiento y los que se tienen que utilizar como test
muestra<-sample(2,n,replace=TRUE, prob = c(0.70,0.30))
table(muestra)  # Añadimos la variable muestra a nuestra base de datos
hogar$muestra<-muestra
#Seleccionamos la muestra de entrenamiento
datos_entrena<- subset(hogar,muestra==1)
datos_entrena<-datos_entrena[,-13]
summary(datos_entrena)
#Calculamos las frecuencias relativas de la variable dependiente
prop.table(table(datos_entrena$policy_status_at_t))
#Seleccionamos la de prueba
datos_test<- subset(hogar,muestra==2)
datos_test<-datos_test[,-13]
(summary(datos_test))
# Definimos una semilla para garantizar que cada vez que ejecutemos siempre vamos a obt
ener las mismas muestras
set.seed(1234)
#Generación de la muestras 70% vs 30% # Contamos el número de casos
n_a<-nrow(auto)
# Definimos una variable que se llama muestra que nos identificaran los casos que se ti
enen que utilizar como entrenamiento y los que se tienen que utilizar como test
muestra_a<-sample(2,n_a,replace=TRUE, prob = c(0.70,0.30))
table(muestra_a)  # Añadimos la variable muestra a nuestra base de datos
auto$muestra<-muestra_a
#Seleccionamos la muestra de entrenamiento
datos_entrena_a<- subset(auto,muestra==1)
datos_entrena_a<-datos_entrena_a[,-23]
summary(datos_entrena_a)
#Calculamos las frecuencias relativas de la variable dependiente
prop.table(table(datos_entrena_a$policy_status_at_t))
#Seleccionamos la de prueba
datos_test_a<- subset(auto,muestra==2)
datos_test_a<-datos_test_a[,-23]
(summary(datos_test_a))


# ----------------------------
# Descriptiva variable: last_premium_paid
# ----------------------------
# Analisis Descriptivo Univariante
```

```
# Procedimentos Graficos Estandard
ggplot(auto, aes(x=last_premium_paid)) +
  geom_histogram(aes(y = ..count..*100/sum(..count..)), fill="darkolivegreen", color="#
e9ecef") +
  scale_x_continuous(name="Ultima poliza pagada") +
  scale_y_continuous(name="%")
ggplot(auto,aes(y=last_premium_paid))+
  geom_boxplot(fill="darkolivegreen", color="#69b3a2")
summary(auto$last_premium_paid)
# ----------------------------
# CATEGORICAS
# ----------------------------
# ----------------------------
# Descriptiva variable: dif_current_previous
# ----------------------------
as.factor(dif_current_previous)
# Procedimentos Graficos Estandard
ggplot(auto) + geom_bar(aes(x = dif_current_previous),fill="darkolivegreen", color="#69
b3a2")+ coord_flip()
# ----------------------------
# TABLAS CRUZADAS
# ----------------------------
# ----------------------------
# Car_number_of_seats_M
# ----------------------------
#Grafica
auto$Car_number_of_seats_M<-as.factor(auto$Car_number_of_seats_M)
ggplot(auto,aes(x=policy_status_at_t,fill=Car_number_of_seats_M))+
  geom_bar(aes(y = ..count..*100/sum(..count..))) +
  scale_fill_manual(values=c("darkolivegreen", "#A8D45D"))+
  coord_flip()  + labs(y="%")
#Tabla
tab_xtab(var.row = auto$Car_number_of_seats_M, var.col = auto$policy_status_at_t, title
 = "Tabla cruzada variable dependiente: estado de la poliza ", show.row.prc = TRUE)


# ----------------------------
# Auto
# ----------------------------

# ----------------------------
# NN
# ----------------------------
#Auto sin pclaims
NnModel20_a = nnet(policy_status_at_t~Age_client+ previous_to_last_premium_paid+ last_premium_p
aid+ dif_current_previous+ dif_current_first+  Client_Seniority+ Policy_numSupplements+Policy_P
aymentMethod_A+car_type_M_Tour+contracted_guarantee_M_THIRD+ Car_power_M+ car_bonus_M +Car_numb
er_of_seats_M +value_of_car_9k+value_of_car_M+totalpol
, data=datos_entrena_a,size=17,maxit=1000,na.action = "na.omit")
NnModel20_a$value
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff04441270351
6c34f1a4684a5/nnet_plot_update.r')
# COMPARAMOS AHORA LA CAPACIDAD PREDICTIVA DE LAS DOS REDES NEURONALES QUE HEMOS ENTRENADO # CO
N 10 NODOS
plot.nnet(NnModel20_a)
# CON LA MUESTRA DE ENTRENAMIENTO
#Predecimos las probabilidades de renovación del contrato P(X1=0) para cada individuo
predEntrena_NN_a<- predict(NnModel20_a, newdata=datos_entrena_a)
# Predecimos que el individuo renueva si la probabilidad predicha es superior a la proporción d
e 0 en la población
predX1_NN_a<-(predEntrena_NN_a>=0.8040911 )
predTest_NN_a<-predict(NnModel20_a, newdata=datos_test_a)
```

```
obs<-as.factor(datos_entrena_a$policy_status_at_t)
pred<-as.numeric(predX1_NN_a)
conf_mat <- confusionMatrix(table(pred, obs))

# Predecimos que el individuo renueva si la probabilidad predicha es superior a la proporción d
e 0 en la población
predX1_TestNN_a<-(predTest_NN_a>=0.8040911 )
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
xx_a<-table(predX1_NN_a,datos_entrena_a$policy_status_at_t)
round(prop.table(table(predX1_NN_a,datos_entrena_a$policy_status_at_t)),4)
table(predX1_TestNN_a,datos_test_a$policy_status_at_t)
round(prop.table(table(predX1_TestNN_a,datos_test_a$policy_status_at_t)),4)
obs_testa<-as.factor(datos_test_a$policy_status_at_t)
pred_testa<-as.numeric(predX1_TestNN_a)
#MATRIZ DE CONFUSION
conf_mat_testa <- confusionMatrix(table(pred_testa, obs_testa))
#ROC
(ROCRpred_nn2a <- prediction(pred_testa,obs_testa))
perf <- performance(ROCRpred_nn2a, "tpr", "fpr")
plot(perf, main="ROC curve")
abline(0,1,col="blue")


#Auto con pclaims

NnModel20_a2 = nnet(policy_status_at_t~Age_client+ previous_to_last_premium_paid+ last_premium_
paid+ dif_current_previous+ dif_current_first+  Client_Seniority+ Policy_numSupplements+Policy_
PaymentMethod_A+car_type_M_Tour+contracted_guarantee_M_THIRD+ Car_power_M+ car_bonus_M +Car_num
ber_of_seats_M +value_of_car_9k+value_of_car_M+totalpol +pclaims
, data=datos_entrena_a,size=17,maxit=1000,na.action = "na.omit")
NnModel20_a2$value
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff04441270351
6c34f1a4684a5/nnet_plot_update.r')
# COMPARAMOS AHORA LA CAPACIDAD PREDICTIVA DE LAS DOS REDES NEURONALES QUE HEMOS ENTRENADO # CO
N 10 NODOS
plot.nnet(NnModel20_a2)
# CON LA MUESTRA DE ENTRENAMIENTO
#Predecimos las probabilidades de renovación del contrato P(X1=0) para cada individuo
predEntrena_NN_a2<- predict(NnModel20_a2, newdata=datos_entrena_a)
# Predecimos que el individuo renueva si la probabilidad predicha es superior a la proporción d
e 0 en la población
predX1_NN_a2<-(predEntrena_NN_a2>=0.8040911 )
predTest_NN_a2<-predict(NnModel20_a2, newdata=datos_test_a)
# Predecimos que el individuo renueva si la probabilidad predicha es superior a la proporción d
e 0 en la población
predX1_TestNN_a2<-(predTest_NN_a2>=0.8040911)
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
xx_a2<-table(predX1_NN_a2,datos_entrena_a$policy_status_at_t)
round(prop.table(table(predX1_NN_a2,datos_entrena_a$policy_status_at_t)),4)
table(predX1_TestNN_a2,datos_test_a$policy_status_at_t)round(prop.table(table(predX1_TestNN_a2,
datos_test_a$policy_status_at_t)),4)
obs_testap<-as.factor(datos_test_a$policy_status_at_t)
pred_testap<-as.numeric(predX1_TestNN_a2)
(conf_mat_testap <- confusionMatrix(table(pred_testap, obs_testap)))
#ROC
ROCRpred_nn2ap <- prediction(pred_testap,obs_testap)
perfap <- performance(ROCRpred_nn2ap, "tpr", "fpr")
plot(perf, main="Comparacion curva ROC con siniestralidad ",col= "green")
legend(x = "bottomright", legend=c("Con pclaims", "sin pclaims"),
       col=c("green", "red"), lty=1:2, cex=0.8)
```

```
abline(0,1,col="blue")
plot(perfap, add= TRUE, col="red")
aucap <- performance(ROCRpred_nn2ap, measure = "auc")
aucap <- aucap@y.values[[1]]
aucap


# ----------------------------
# SVM
# ----------------------------

load("C:/Users/34644/OneDrive - Universitat de Barcelona/0. TFG/preprosesing.RData")
#Auto sin pclaims
svmmodel_sig_a<-svm(policy_status_at_t~Age_client+ previous_to_last_premium_paid+ last_
premium_paid+ dif_current_previous+ dif_current_first+  Client_Seniority+ Policy_numSup
plements+Policy_PaymentMethod_A+car_type_M_Tour+contracted_guarantee_M_THIRD+ Car_power
_M+ car_bonus_M +Car_number_of_seats_M +value_of_car_9k+value_of_car_M+totalpol,  data=
datos_entrena_a,kernel="sigmoid",na.action = na.omit )
svmmodel_sig_a

# CON LA MUESTRA DE ENTRENAMIENTO #Predecimos directamente la clase de pertenencia de c
ada individuo
predX1_SVM_a<- predict(svmmodel_sig_a, newdata=datos_entrena_a,na.action = na.omit)

# CON LA MUESTRA DE TEST
#Predecimos directamente la clase de pertenencia de cada individuo
predX1_TestSVM_a<-predict(svmmodel_sig_a, newdata=datos_test_a)
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
table(predX1_SVM_a,datos_entrena_a$policy_status_at_t)
prop.table(table(predX1_SVM_a,datos_entrena_a$policy_status_at_t))
# PRUEBA
table(predX1_TestSVM_a,datos_test_a$policy_status_at_t)
prop.table(table(predX1_TestSVM_a,datos_test_a$policy_status_at_t))
# Representación del SMV
plot(svmmodel_sig_a,datos_entrena_a, Age_client ~ Client_Seniority)

predX1_TestSVM_a_<-predict(svmmodel_sig_a,type="prob", newdata=datos_test, probability=
 TRUE)
#MATRIZ DE CONFUSIONconf_mat_test_SVM_a <- confusionMatrix(table(predX1_TestSVM_a, dato
s_test_a$policy_status_at_t))
plot(svmmodel_sig_a,datos_entrena_, Client_Seniority~ last_premium_paid)


#Auto con pclaims

svmmodel_sig_a2<-svm(policy_status_at_t~Age_client+ previous_to_last_premium_paid+ last
_premium_paid+ dif_current_previous+ dif_current_first+  Client_Seniority+ Policy_numSu
pplements+Policy_PaymentMethod_A+car_type_M_Tour+contracted_guarantee_M_THIRD+ Car_powe
r_M+ car_bonus_M +Car_number_of_seats_M +value_of_car_9k+value_of_car_M+totalpol+pclaim
s,  data=datos_entrena_a,kernel="sigmoid",na.action = na.omit )

svmmodel_sig_a2
# CON LA MUESTRA DE ENTRENAMIENTO #Predecimos directamente la clase de pertenencia de c
ada individuo
predX1_SVM_a2<- predict(svmmodel_sig_a2, newdata=datos_entrena_a,na.action = na.omit)

# CON LA MUESTRA DE TEST
#Predecimos directamente la clase de pertenencia de cada individuo
predX1_TestSVM_a2<-predict(svmmodel_sig_a2, newdata=datos_test_a)
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
table(predX1_SVM_a2,datos_entrena_a$policy_status_at_t)
prop.table(table(predX1_SVM_a2,datos_entrena_a$policy_status_at_t))
```

```
# PRUEBA
table(predX1_TestSVM_a2,datos_test_a$policy_status_at_t)
prop.table(table(predX1_TestSVM_a2,datos_test_a$policy_status_at_t))
# Representación del SMV
plot(svmmodel_sig_a2,datos_entrena_a, Age_client ~ Client_Seniority)

obs_test_SVM_a2<-as.numeric(as.character(datos_test_a$policy_status_at_t))
pred_test_SVM_a2<-as.numeric(as.character(predX1_TestSVM_a2))

#MATRIZ DE CONFUSION
conf_mat_test_SVM_a2 <- confusionMatrix(table(pred_test_SVM_a2, obs_test_SVM_a2))
# Representación del SMV
plot(svmmodel_sig_a2,datos_entrena, Client_Seniority~ last_premium_paid)

# ----------------------------
# Hogar
# ----------------------------

# ----------------------------
# NN
# ----------------------------
#HOGAR sin pclaims
NnModel1_h2 = nnet(policy_status_at_t~., data=datos_entrena,size=6,maxit=10000,na.actio
n = "na.omit")
# Error
NnModel1_h2$value
NnModel20_h = nnet(policy_status_at_t~ sex_customer+Age_client+Client_Seniority+Policy_
PaymentMethod_A+dif_current_previous + dif_current_first+Insuredcapital_content_H+Insur
edcapital_continent_H+totalpol+Policy_PaymentMethod_A+HomeType_H_PI,  data=datos_entren
a ,size=6, maxit=10000,na.action = "na.omit")
NnModel20_h$value
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044
412703516c34f1a4684a5/nnet_plot_update.r')
# COMPARAMOS AHORA LA CAPACIDAD PREDICTIVA DE LAS DOS REDES NEURONALES QUE HEMOS ENTREN
ADO # CON 10 NODOS
plot.nnet(NnModel20_h)
# CON LA MUESTRA DE ENTRENAMIENTO
#Predecimos las probabilidades de renovación del contrato P(X1=0) para cada individuo
predEntrena_NN_h<- predict(NnModel20_h, newdata=datos_entrena)
# Predecimos que el individuo renueva si la probabilidad predicha es superior a la prop
orción de 0 en la población
predX1_NN_h<-(predEntrena_NN_h>=0.8901545 )
predTest_NN_h<-predict(NnModel20_h, newdata=datos_test)

# Predecimos que el individuo renueva si la probabilidad predicha es superior a la prop
orción de 0 en la población
predX1_TestNN_h<-(predTest_NN_h>=0.8901545 )
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
xx_h<-table(predX1_NN_h,datos_entrena$policy_status_at_t)
round(prop.table(table(predX1_NN_h,datos_entrena$policy_status_at_t)),4)
table(predX1_TestNN_h,datos_test$policy_status_at_t)
round(prop.table(table(predX1_TestNN_h,datos_test$policy_status_at_t)),4)
obs_test_h<-as.factor(datos_test$policy_status_at_t)
pred_test_h<-as.numeric(predX1_TestNN_h)
#MATRIZ DE CONFUSION
(conf_mat_test_h <- confusionMatrix(table(pred_test_h, obs_test_h)))
#ROC
ROCRpred_nn2_h <- prediction(pred_test_h,obs_test_h)
perf_h <- performance(ROCRpred_nn2_h, "tpr", "fpr")
plot(perf_h, main="ROC curve")
abline(0,1,col="blue")
auc_h <- performance(ROCRpred_nn2_h, measure = "auc")
```

```
auc_h<- auc_h@y.values[[1]]
auc_h

#HOGAR con pclaims

NnModel20_h2 = nnet(policy_status_at_t~., data=datos_entrena,size=6,maxit=10000,na.acti
on = "na.omit")
# Error
NnModel20_h2$value
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044
412703516c34f1a4684a5/nnet_plot_update.r')
# COMPARAMOS AHORA LA CAPACIDAD PREDICTIVA DE LAS DOS REDES NEURONALES QUE HEMOS ENTREN
ADO # CON 10 NODOS
plot.nnet(NnModel20_h2)
# CON LA MUESTRA DE ENTRENAMIENTO
#Predecimos las probabilidades de renovación del contrato P(X1=0) para cada individuo
predEntrena_NN_h2<- predict(NnModel20_h2, newdata=datos_entrena)
# Predecimos que el individuo renueva si la probabilidad predicha es superior a la prop
orción de 0 en la población
predX1_NN_h2<-(predEntrena_NN_h2>=0.8901545 )
predTest_NN_h2<-predict(NnModel20_h2, newdata=datos_test)

# Predecimos que el individuo renueva si la probabilidad predicha es superior a la prop
orción de 0 en la población
predX1_TestNN_h2<-(predTest_NN_h2>=0.8901545 )
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
xx_h2<-table(predX1_NN_h2,datos_entrena$policy_status_at_t)
round(prop.table(table(predX1_NN_h2,datos_entrena$policy_status_at_t)),4)
table(predX1_TestNN_h2,datos_test$policy_status_at_t)
round(prop.table(table(predX1_TestNN_h2,datos_test$policy_status_at_t)),4)
obs_test_h2<-as.factor(datos_test$policy_status_at_t)
pred_test_h2<-as.numeric(predX1_TestNN_h2)
#MATRIZ DE CONFUSION
(conf_mat_test_h2 <- confusionMatrix(table(pred_test_h2, obs_test_h2)))
#ROC
ROCRpred_nn2_h2 <- prediction(pred_test_h2,obs_test_h2)
perf_h2 <- performance(ROCRpred_nn2_h2, "tpr", "fpr")


plot(perf_h2, main="Comparacion curva ROC con siniestralidad ",col= "green")
legend(x = "bottomright", legend=c("Con pclaims", "sin pclaims"),
       col=c("green", "red"), lty=1:2, cex=0.8)
abline(0,1,col="blue")
plot(perf_h, add= TRUE, col="red")


# ---------------------------
# SVM
# ---------------------------


#HOGAR sin pclaims
svmmodel_sig_h<-svm(policy_status_at_t~sex_customer+Age_client+Client_Seniority+Policy_
PaymentMethod_A+dif_current_previous +
dif_current_first+Insuredcapital_content_H+Insuredcapital_continent_H+
                      totalpol+Policy_PaymentMethod_A+HomeType_H_PI , data=datos_entren
a,kernel="sigmoid",na.action = na.omit )
svmmodel_sig_h
# CON LA MUESTRA DE ENTRENAMIENTO #Predecimos directamente la clase de pertenencia de c
ada individuo
predX1_SVM_h<- predict(svmmodel_sig_h, newdata=datos_entrena,na.action = na.omit)

# CON LA MUESTRA DE TEST
```

```
#Predecimos directamente la clase de pertenencia de cada individuo
predX1_TestSVM_h<-predict(svmmodel_sig_h, newdata=datos_test)
#MATRICES DE CONFUSIÓN PARA LAS MUESTRAS DE:
# ENTRENAMIENTO
table(predX1_SVM_h,datos_entrena$policy_status_at_t)
prop.table(table(predX1_SVM_h,datos_entrena$policy_status_at_t))
# PRUEBA
table(predX1_TestSVM_h,datos_test$policy_status_at_t)
prop.table(table(predX1_TestSVM_h,datos_test$policy_status_at_t))
# Representación del SMV
plot(svmmodel_sig_h,datos_entrena, Age_client ~ Client_Seniority)


#HOGAR con pclaims

predX1_TestSVM_h_<-predict(svmmodel_sig_h,type="prob", newdata=datos_test, probability=
 TRUE)
#MATRIZ DE CONFUSION
(conf_mat_test_SVM_h <- confusionMatrix(table(predX1_TestSVM_h, datos_test$policy_statu
s_at_t)))
plot(svmmodel_sig_h,datos_entrena, Client_Seniority~ Insuredcapital_continent_H)
```