Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

Título: La importancia de los modelos de retención de clientes en las entidades aseguradoras

Autor: Albert Romero Martínez

Tutora: Catalina Bolancé

Curso académico: 2017 - 2018

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Final de Máster

Máster en Ciencias Actuariales y Financieras

La importancia de los modelos de retención de clientes en las entidades aseguradoras

Autor: Albert Romero Martínez

Tutora: Catalina Bolancé

“El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto”

Resumen

En un entorno macroeconómico global cada vez más competitivo, captar nuevos clientes y retenerlos es el gran reto del siglo XXI. Los consumidores están cada vez más informados, tienen un amplio abanico de posibilidades y tienen un gran poder de decisión. ¿Cómo captar su interés? Esta pregunta es la que se hacen la mayoría de las entidades aseguradoras, puesto que esta gran competitividad ha hecho que éstas se den cuenta que su bien más preciado no es otro que el cliente, evidenciándose la gran necesidad de retención del mismo en sus carteras. Para poder identificar cuáles son las necesidades de los clientes y qué aspectos hacen variar la decisión de renovación o anulación de la póliza en las compañías de seguros, se necesitan herramientas que permitan analizarlo.

Los modelos lineales generalizados son modelos estadísticos que se aplican para diferentes usos en el campo actuarial, como por ejemplo, para realizar modelos de tarificación. Sin embargo, la tarificación no es la única aplicación interesante de estos modelos, pues también cabe destacar la importancia de los modelos de retención de clientes. El texto aquí comprendido analiza dos tipos de modelos GLM - univariantes y bivariantes - aplicables a una muestra real de una cartera con dos tipos de seguros diferentes, así como también se comprueba cuál de los dos se ajusta mejor, para descubrir por qué son necesarios hoy en día los modelos de retención de clientes y qué nos pueden aportar.

Palabras clave: retención de clientes, modelos lineales generalizados, criterio de Akaike, seguro de hogar, seguro de automóvil, estimación, tabla de clasificación.

Abstract

In an increasingly competitive global macroeconomic environment, attracting new customers and retaining them is the biggest challenge of the 21st century. Customers are increasingly informed, as they have a wide range of possibilities and a great power of decision. How should companies capture their interest? This is a recurrent question for insurance companies, since this great competitiveness has made them realize that their most precious asset is none other than the customer, evidencing the great need to retain them in their portfolios. In order to identify which are the needs of the customers and which aspects make them vary their decision to renew or cancel the insurance policy, insurance companies must have tools to analyse it.

The generalized linear models are statistical models that are applied for different uses in the actuarial field, such as for making pricing models. Nevertheless, pricing is not the only interesting application for these models, as it is also worth to note the importance of customer retention models. The text shown here analyses two types of GLM models - univariate and bivariate - applicable to a real sample of a portfolio with two different types of insurances, as it is also checked which of them fits better, to discover why customer retention models are necessary nowadays and how can help us retaining customers.

Key words: customer retention, generalized linear models, Akaike criterion, home insurance, motor insurance, estimation, classification table.

Índice

Listado de Gráficos

Listado de Tablas

1) Introducción; presentación y objetivos del trabajo

Este trabajo se fundamenta en la investigación y el estudio de la importancia de los modelos de retención de clientes para las entidades aseguradoras.

El primer planteamiento que se le puede ocurrir al lector del presente trabajo puede ser ¿por qué centrar el trabajo en la modelización de modelos estadísticos para cuantificar la probabilidad de retención de clientes en una cartera de una entidad aseguradora?

Los modelos de retención de clientes se utilizan en el sector asegurador para prever qué cantidad de clientes de la cartera abandonarían la misma cuando se hacen cambios en algunas de las variables que la determinan. Por ejemplo, si una empresa estima oportuno el incremento generalizado de las primas a partir del 2018, necesita cuantificar de una manera más o menos exacta, con los factores de riesgo y los datos de los que dispone, cómo van a actuar sus clientes y, por tanto, cuántos de ellos no van a aceptar dicho incremento.

Esta apreciación se hace especialmente necesaria para entidades con un gran volumen de negocio y, por tanto, de clientes, ya que la complejidad que supone la elaboración de modelos de retención de clientes para cada producto del negocio asegurador de la empresa y la inversión que supone en términos de personal y tiempo, puede suponer más un inconveniente que una ventaja para empresas de tamaño más reducido.

Adicionalmente, los modelos de retención tienen un interés añadido, y es que permiten apreciar cómo afecta al comportamiento del cliente el hecho de tener contratada una simple póliza de un producto, o bien tener más de una a la vez.

Un apunte conceptual interesante a matizar es que en este trabajo se modelizará la retención de clientes en una cartera no vida. Como es sabido, existen dos ramas en el mundo de los seguros; la rama de los seguros no vida y la de los seguros de vida.

La primera rama engloba los seguros relacionados con las coberturas de siniestros a corto plazo, es decir, normalmente con un intervalo temporal de cobertura anual, y los que el capital a pagar viene determinado por la siniestralidad observada. Un ejemplo de este tipo de seguros podría ser el seguro de automóvil, el cual cubre principalmente la responsabilidad civil derivada de los daños realizados a terceros por el asegurado con su vehículo, aparte de poder incluir otro tipo de coberturas adicionales.

La segunda rama engloba los seguros relacionadas con las coberturas de siniestros a largo plazo, es decir, normalmente con un intervalo temporal de cobertura mayor que uno o varios años, y en los que la prestación a pagar viene determinada des del momento en que se formaliza el contrato.

Un ejemplo de este tipo de seguros es el seguro de fallecimiento, en el que se asegura el capital que se pagará a los beneficiarios cuando el asegurado fallezca (dentro del período de cobertura).

Así pues, la cartera que se va a usar para modelizar la retención de clientes va a ser una muestra extraída de una cartera de seguros no vida con productos de automóvil y de hogar, formada por 10.041 registros, sobre los que se miden 18 variables.

Por otro lado, la decisión de realizar este trabajo ha sido motivada por diferentes razones. La primera surgió de la curiosidad, ya que durante un periodo aproximado de 10 meses realicé un periodo de prácticas en una gran multinacional del mundo de los seguros, concretamente en la oficina de cálculo actuarial no vida, viendo de cerca los procesos de modelización que se llevaban a cabo. Allí es donde descubrí que aparte de los modelos de tarificación que se utilizan para calcular las primas adecuadas a los seguros de no vida, los modelos estadísticos también se podían utilizar para calcular la retención del cliente o, como allí lo llamaban, para modelizar las anulaciones.

La segunda razón fue que posteriormente en la universidad se nos planteó la posibilidad de realizar un "business case" en colaboración con otra gran empresa multinacional del sector de los seguros, lo cual me permitió profundizar un poquito más en cuál era la estrategia de una compañía aseguradora para captar al cliente, dependiendo de si la compañía era mediada o no, entre otros aspectos. Fue un enfoque menos técnico que el que el lector encontrará en este texto, pero fue muy valioso para entender el funcionamiento de los pasos a seguir durante la renovación de las pólizas de los clientes, además de ver qué factores de riesgo podían afectar en sus decisiones. Al acabar este "business case", se hicieron las presentaciones y mi equipo tuvo la suerte de ser el ganador.

La tercera y última razón ha surgido a partir de las dos anteriores, ya que la curiosidad que el "business case" me despertó hacía los procesos de retención de clientes junto con la importancia que he visto que tienen los modelos estadísticos durante la realización de las prácticas han hecho que quiera descubrir la verdadera importancia que tienen los modelos de retención de clientes en una entidad aseguradora. Es decir, ver y analizar los resultados obtenidos mediante la aplicación de los modelos estadísticos que expondré a continuación con el objetivo de entender si dichos modelos deberían ser imprescindibles para la mayoría de las entidades aseguradoras.

Los modelos que se van a utilizar son tres; dos modelos univariantes y un modelo bivariante. Los modelos univariantes son aquellos que se centran en modelizar una única variable, estos son el "logit univariante" y el "probit univariante". En cambio, los modelos bivariantes son aquellos que se usan para modelizar la distribución conjunta de dos variables, este es el "probit bivariante".

A la luz de los resultados obtenidos, se compararán los dos modelos univariantes para ver cuál es el que nos da un mejor resultado y posteriormente analizar las diferencias que surgen respecto al modelo bivariante. Se ha escogido estos dos tipos de modelo debido a que contamos con una base de datos con dos variables dependientes y, por tanto, para los objetivos del presente texto los modelos univariantes y bivariantes son suficiente.

Así pues, a partir de lo comentado anteriormente, el trabajo se fundamenta en el planteamiento de una serie de cuestiones.

- ¿Cuál es el papel de los modelos de retención en las compañías aseguradoras?
- ¿Por qué surge la necesidad de modelizar este tipo de modelos?
- ¿Qué modelos pueden ser útiles para la modelización de la retención de clientes en una empresa aseguradora?
- ¿Partiendo de una cartera de asegurados con dos tipos de producto, afecta en el comportamiento del cliente el tener una o más pólizas con la compañía?
- ¿Si es así, qué diferencias pueden surgir de la aplicación de diferentes tipos de modelos?
- ¿Cuál es el modelo que mejor nos explica dicho comportamiento?

De todas estas preguntas claves, se derivan los objetivos fundamentales del trabajo, que son el estudio y comprensión de los principales tipos de modelos estadísticos aplicables para la modelización de la retención de los clientes en cartera, el análisis de las variables que afectan en un modelo de retención de clientes, el contraste de los resultados obtenidos al utilizar dichos modelos estadísticos y, por último, la determinación del modelo que mejor explica el comportamiento de los clientes.

Para poder realizar este trabajo ha sido necesario la búsqueda de referencias bibliográficas que contextualicen la metodología utilizada, así como su uso concreto. En el marco conceptual más general podemos encontrar referencias a los modelos utilizados en este texto en Cameron y Trivedi (2005), Takeshi (1985) o en Greene (1999), siendo este tercero la bibliografía de referencia usada en el apartado en que se explica la metodología empleada.

En un ámbito de aplicación de dichos modelos en campos actuariales, encontramos referencias como Frees (2009) o Frees et al. (2014), en este último podemos encontrar un interesante apartado dedicado a las regresiones con variables dependientes categóricas.

A modo de revisión más concreta de los posibles usos que ofrecen los modelos GLM, encontramos ejemplos como en el trabajo de Guillen et al. (2012), en el que se mide los efectos que varían con el tiempo en el estudio de la lealtad de los clientes en el ámbito asegurador, o como en Wilson (2009), que utiliza la regresión logística para detectar el fraude en el sector asegurador.

Por último, también podemos encontrar referencias que hayan utilizado modelizaciones GLM para predecir o controlar la retención de clientes en el sector asegurador, en publicaciones tales como Bolancé et al. (2016a), en que para un caso real del sector del asegurador del motor, se usan tres métodos de ajuste diferentes para analizar los resultados de diferentes metodologías y concluir que dependiendo del objetivo del usuario, puede ser óptimo un método u otro, o bien en Bolancé et al. (2016b), con un objetivo parecido al de la publicación anterior y en el que también se aplican dichos métodos en los seguros de hogar y automóviles, como sucede en el presente texto.

Así pues, el texto aquí expuesto se estructura en tres grandes apartados acompañados de la presente introducción y de las conclusiones finales obtenidas a partir de la realización de este texto.

El apartado más teórico se basa en la explicación de la metodología empleada, es decir, se ofrece una explicación detallada de los modelos que se van a usar posteriormente y se explica en qué consiste el criterio de información de Akaike usado para determinar qué modelo es el que mejor se ajusta.

Por otra parte, en el siguiente apartado se detalla en qué consiste la base de datos sobre la que se aplicarán los modelos, así como también se explica la distribución de las variables que forman parte de la misma.

El tercero y último apartado, es en el que se lleva a cabo la aplicación práctica de los conceptos descritos con anterioridad. En esta parte se tratan los datos y se les aplican los modelos, se obtienen los resultados y se analizan con profundidad las modelizaciones efectuadas para obtener las conclusiones lo más precisas posibles.

Además, se determina cuál es el mejor modelo de retención de clientes de los probados, cuál es la tasa de retención de cada seguro y se comparan los resultados de las estimaciones efectuadas - con las tasas de retención obtenidas con distintos puntos de corte - con los datos reales.

2) Metodología; introducción a los modelos aplicados

El objetivo de este trabajo es la determinación de la tasa de retención sobre una cartera concreta y, por tanto, analizar la importancia de los modelos de retención de clientes para las entidades aseguradoras.

Las variables dependientes o explicadas, que en nuestro caso nos indican si un cliente tiene la póliza de automóvil o de hogar en vigor o anulada, son variables que tienen dos posibles respuestas y que, por tanto, conllevan una elección cualitativa y binaria por parte del tomador: renovar la póliza o anularla. Además, al ser variables discretas, no permiten la aplicación de modelos de regresión clásicos.

Estas variables dependientes vienen condicionadas por la influencia de las variables independientes o explicativas, la confluencia de las cuales infiere en cada cliente de manera desigual haciendo que éste tome una decisión u otra.

Como ya se ha comentado anteriormente, la variable dependiente es una variable binomial y, por tanto, para saber qué probabilidad de que renueve existe no se puede utilizar un modelo de regresión, sino que se debe aplicar un modelo de elección binaria (aunque al fin y al cabo el objetivo de ambos modelos sea la misma). Tal y como se explica en Greene (1999), *"a partir de datos de la variable de interés y de un conjunto de variables explicativas, el analista quiere especificar una relación entre una y las otras, más o menos de modo similar como se ha hace con los modelos de regresión"*.

Como los modelos que se explican a continuación son modelos no lineales, ya que la relación entre la esperanza de las variables dependientes y las variables explicativas no es lineal, el método de estimación utilizado es el de máximo verosimilitud.

2.1) Modelos univariantes

En este texto se tratan dos tipos de modelos univariantes: el modelo probit univariante y el modelo logit univariante.

Al ser modelos univariantes, no interrelacionarán el efecto que pueda tener sobre la renovación de un seguro el hecho de que un asegurado renueve el otro seguro. Es decir, si un asegurado, dados unos factores de riesgo favorables, elige renovar el seguro de hogar, no afectará en su decisión de renovar o no el seguro de automóvil.

Esto se traduce en una modelización de la retención para los seguros de hogar y los seguros de automóvil por separado, es decir, en base a las variables explicativas de cada seguro se obtendrán unos resultados a partir de los que explicar el comportamiento de los asegurados y predecir que va a suceder.

Entrando en materia, a partir de lo dispuesto en Greene (1999), podemos decir que, en el período de renovación de las pólizas, los asegurados pueden decidir entre renovar el seguro (Y=1) y, por tanto, continuar disfrutando el año siguiente de la cobertura ofrecida por su compañía a cambio de una contraprestación en forma de prima, o bien anular el seguro (Y=0). Esta decisión puede venir dada por un conjunto de factores tales como la antigüedad de la póliza, la edad del asegurado u otros factores particulares relacionados con los seguros de hogar y automóviles, los cuales quedan recogidos en un vector fila "x" con un primer término igual a uno y con "p" variables. Además, el vector columna de "p" parámetros "β" refleja la magnitud del impacto que tiene cada factor de riesgo sobre la probabilidad.

De esta manera, podemos definir la probabilidad de que ocurra un suceso u otro como:

$$Prob(Y = 1) = F(x \cdot \beta) \qquad Prob(Y = 0) = 1 - F(x \cdot \beta)$$

siendo la probabilidad de Y=1 el suceso que modelizaremos.

Como se ha comentado anteriormente, utilizar un modelo de probabilidad lineal, tal como un modelo de regresión, no sería la opción más adecuada, ya que nos provocaría varios inconvenientes al no haber linealidad en la relación entre la variable dependiente y las explicativas. La no linealidad implica que no se pueden interpretar los efectos marginales de cada factor, ya que no son constantes. Por tanto, sólo se puede interpretar su signo, exceptuando los casos en que la escala de las variables explicativas es la misma, en los cuales también se pueden comparar sus magnitudes.

Además tal y como se especifica en Greene (1999), *"no podemos asegurar que las predicciones de este modelo (refiriéndose al modelo lineal) parezcan verdaderas probabilidades, a menos que hagamos manipulaciones específicas para que así sea. No podemos restringir "β'x" (entendiendo que F(x·β) = β'x) al intervalo [0,1]. Ello origina tanto varianzas negativas como probabilidades imposibles. Por esta razón, […] el modelo lineal se utiliza cada vez menos […]. Lo que queremos es, por tanto, un modelo que proporcione predicciones consistentes a la teoría que subyace a"*:

$$Prob(Y = j) = F[efectos\ relevantes: parámetros]$$

siendo "j" el suceso que ocurre.

Por tanto, para un vector de regresores dados esperaríamos que:

$$\lim_{\beta' x \to +\infty} Prob(Y = 1) = 1 \qquad \text{y} \qquad \lim_{\beta' x \to -\infty} Prob(Y = 1) = 0$$

Para solucionar este problema, basta con aplicar una función de distribución no negativa definida sobre la recta real; siendo las distribuciones utilizadas más comunes la distribución normal estándar y la distribución logística.

Al utilizar la distribución normal se da lugar al modelo probit univariante, que se especifica como:

$$Prob(Y = 1) = \int_{-\infty}^{\beta' x} \phi(t)dt = \Phi(\beta' x),$$

donde "$\Phi(\beta' x)$" es la función de distribución normal estándar de la primera derivada de los parámetros "β" por el vector de factores de riesgo "x".

Por otro lado, si la función que se utiliza es la distribución logística, el modelo que obtenemos es el modelo logit univariante, que se especifica como:

$$Prob(Y = 1) = \frac{e^{\beta' x}}{1+e^{\beta' x}} = \Lambda(\beta' x),$$

donde "$\Lambda(\beta' x)$" es la función de distribución logística de la primera derivada de los parámetros "β" por el vector de factores de riesgo "x".

La semejanza entre ambas distribuciones es notable, excepto en cuanto se trata de las colas, pues en el caso del modelo logit éstas son más pesadas. Es por eso que, tal y como se expone en Greene (1999), *"las dos distribuciones tienden a dar probabilidades muy similares a los valores intermedios de "β'x" [...]. La distribución logística tiende a dar probabilidades mayores que la distribución normal al suceso Y = 0 (que no se renueve la póliza) cuando "β'x" es muy pequeño (y probabilidades menores que la distribución normal al suceso Y = 0 cuando "β'x" es muy grande). Pero para poder extraer de aquí una regla general, sería necesario conocer el valor de β"*.

Es por lo dispuesto anteriormente, que predeciblemente los dos modelos anteriormente citados nos proporcionaran predicciones diferentes si la muestra según Greene (1999), contiene:

I. Pocas respuestas afirmativas (valores de Y = 1) o pocas respuesta negativas (valores de Y = 0).
II. Gran variación en una variable independiente de importancia, especialmente si también se cumple I.

A priori, nuestra cartera contiene muchas renovaciones y muy pocas anulaciones, tanto en los seguros de hogar como en los de automóvil, por lo que podría ser que ambos modelos nos abocaran a predicciones algo diferentes. En el apartado práctico veremos y comentaremos los resultados obtenidos con ambos modelos.

2.2) Modelos bivariantes

En la realidad, que un asegurado decida renovar o no un tipo de seguro dentro de una compañía sí que puede tener efectos sobre la decisión a tomar en el otro tipo de seguro.

Pongamos un ejemplo: un asegurado de cincuenta años quiere renovar el seguro de automóvil con nuestra compañía porque este año no ha tenido siniestralidad y el precio no le ha subido. Además, se muestra conforme con las coberturas que ofrece el seguro y cree que la asistencia que le han proporcionado otros años en los que sí ha tenido siniestros ha sido buena. Por tanto, este seguro de automóvil le da seguridad y comodidad a nuestro asegurado. Por otro lado, este año el seguro de hogar que tiene contratado, también con nuestra compañía, le ha incrementado un poco el precio, sin que él haya tenido siniestralidad. La compañía ha tenido que hacer unos retoques en algunos coeficientes de algunos factores de riesgo porque han detectado que no cubrían bien el riesgo de dichos factores. El asegurado se muestra un poco descontento con este incremento y, aunque tiene alguna oferta de la competencia con una prima a pagar inferior, decide renovar también con nosotros. Ese hecho puede venir influenciado por tener otro seguro con la compañía; ya que quizás el asegurado le ha dado importancia al hecho de tener ambos seguros en una misma compañía, o bien, ha tenido en cuenta el buen servicio prestado en el seguro de automóviles otros años y piensa que si ocurre un siniestro en el inmueble estará bien cubierto, o simplemente por la comodidad de no tener que estar pendiente de controlar varios seguros en diferentes compañías con las que aún no ha tenido experiencia.

Es esta posible interrelación entre variables dependientes la que nos apremia a aplicar otro tipo de modelos; los modelos bivariantes.

Estos modelos, a diferencia de los primeros, sí que tienen en cuenta la interrelación que una variable dependiente pueda ejercer sobre la otra y, por tanto, recogen el efecto de una decisión de renovación o anulación de un asegurado sobre un tipo de seguro sobre la decisión de renovar o anular el otro tipo de seguro.

Así pues, en este texto se usa un tipo de modelo bivariante; el probit bivariante.

En el probit bivariante tenemos dos ecuaciones de utilidad (una para cada tipo de seguro; hogar y automóviles) que son las ecuaciones que nos indicarán si ocurrirá el suceso "renovación" $Y_1 = 1$ en el caso de los seguros de hogar e $Y_2 = 1$ en el caso de los seguros de automóvil. Las dependencias entre ambas ecuaciones se dan cuando los errores de las mismas están correlacionados.

La especificación del modelo probit bivariante con dos ecuaciones correspondientes a los seguros de hogar y los seguros de automóvil es la siguiente (Greene, 1999):

$$Y_1^* = \beta_1' x_1 + \epsilon_1, \quad Y_1 = 1 \ si \ Y_1^* > 0, \text{ó } 0 \ en \ caso \ contrario,$$

$$Y_2^* = \beta_2' x_2 + \epsilon_2, \quad Y_2 = 1 \ si \ Y_2^* > 0, \text{ó } 0 \ en \ caso \ contrario,$$

$$E[\epsilon_1] = E[\epsilon_2] = 0,$$

$$Var[\epsilon_1] = Var[\epsilon_2] = 1,$$

$$Cov[\epsilon_1, \epsilon_2] = \rho.$$

siendo i = 1,2, dónde 1 = seguro de hogar y 2 = seguro de automóvil:

Y_i^*: Ecuación de utilidad del seguro i=1,2,

β_i': Derivada primera del vector de parámetros "β" del seguro i=1,2,

x_i: Vector de factores de riesgo "x" del seguro i=1,2,

ϵ_i: Término de error de la ecuación de utilidad del seguro i=1,2,

$E[\epsilon_i]$: Esperanza del término de error del seguro i=1,2,

$Var[\epsilon_i]$: Varianza del término de error del seguro i=1,2,

$Cov[\epsilon_1, \epsilon_2]$: Covarianza de los términos de error de los seguros de hogar y automóvil.

Como podemos comprobar en las formulaciones anteriores, siempre que la utilidad de un seguro sea positiva, el asegurado renovará otro año más, ya que dicha renovación le reportará una utilidad superior que si no renovara. Además, podemos afirmar que el modelo probit bivariante presenta homocedasticidad, ya que la varianza de los errores se mantiene constante a lo largo de las observaciones (Greene, 1999).

Por otro lado, es muy interesante tener en cuenta el término "ρ", el cual nos indica la covarianza entre los errores de las dos ecuaciones. Tal y como se indica en Greene (1999), si la covarianza de los errores condicionales de los dos seguros es cero, significa que la covarianza es nula, y que por tanto el modelo está formado por dos ecuaciones probit independientes que pueden ser estimadas por separado, o lo que es lo mismo, el caso univariante anteriormente estudiado. Por lo contrario, si la covarianza no es 0 quiere decir que existe correlaciones entre los términos de error y, por tanto, entre las ecuaciones.

Para poder comprobar si existe o no correlación entre ambas ecuaciones, se debe realizar un contraste de correlación (el cual se realizará en el apartado práctico) con las siguientes hipótesis:

$$H_0: \rho = 0\ (hipótesis\ nula) \qquad H_A: \rho \neq 0\ (hipótesis\ alternativa)$$

En el caso que no se rechace la hipótesis nula de que la covarianza es igual a cero, no existirá correlación entre ambas ecuaciones de utilidad. En caso contrario, existirá correlación entre las ecuaciones de utilidad de los seguros de hogar y automóviles.

Según lo dispuesto en Greene (1999), la función de distribución normal bivariante del modelo probit bivariante es:

$$Prob(x_1 \cdot \beta_1 \leq z_1, x_2 \cdot \beta_2 \leq z_2) = \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \phi_2(t_1, t_2, \rho) dt_1 dt_2 \ = \ \Phi_2(z_1, z_2, \rho)$$

dónde "$\Phi_2(z_1, z_2, \rho)$" es la función de distribución normal bivariante del vector de factores de riesgo de los seguros de hogar "z_1", del vector de factores de riesgo de los seguros de automóvil "z_2" y del término de covarianza entre los errores de ambos seguros "ρ".

La interpretación de los parámetros "β" se realiza del mismo modo que en el caso univariante. Aplicando este razonamiento para las dos ecuaciones de nuestro modelo, obtendremos que la probabilidad bivariante a partir de la cual analizar los posibles efectos marginales que sucedan (Greene, 1999):

$$Prob[y_1 = 1, y_2 = 1] = \Phi_2\ [\beta_1' x, \beta_2' x, \rho]$$

2.3) El criterio de información de Akaike

Una vez la base de datos sea modelizada con modelos univariantes y bivariantes y se hayan analizado los signos de los estimadores de los parámetros y su significación, se planteará la duda de qué modelo es el más fiable o el que nos ofrece unos resultados más ajustados a la realidad.

Para poder dirimir cual es el mejor modelo se suelen usar medidas de bondad de ajuste, las cuales nos proporcionan un indicador de la calidad predictiva de cada modelo.

La medida más utilizada para comprar diferentes modelos GLM acostumbra a ser el criterio de información de Akaike, ampliamente conocido por sus siglas AIC. Dicha medida nos proporciona una cifra para cada modelo que nos permite saber, si se comparan éstos, que la que sea menor indica que dicho modelo es el que mejor se ajusta a los datos reales. El criterio de información de Akaike no nos proporciona información acerca de la calidad del modelo en valores absolutos, sino que es una medida que se usa en valor relativo en comparación al resto de modelos utilizados.

Así, nos permite elegir el modelo que mejor se ajusta y, por tanto, con el que se debe de trabajar si se quieren conseguir unos datos lo más fiables posibles a partir de los que poder sacar conclusiones.

En este caso particular, se plantea un problema adicional al intentar seleccionar el modelo que mejor se ajusta; no podemos comparar los AIC que nos proporcionan los modelos univariantes con el AIC del modelo bivariante, puesto que en un caso se modeliza una variable dependiente y en el otro caso dos.

Para poder llevar a cabo la comparativa de los modelos es necesario juntar los modelos probit univariante de hogar y automóviles, así como también los modelos logit univariante de hogar y automóviles mediante la siguiente formulación:

$$AIC = 2 * k - 2 * logL$$

En el caso concreto que nos concierne, la formulación acaba siendo la siguiente:

$$AIC = 2 * \left(\sum_{i=1}^{2} k_i\right) + \left(\sum_{i=1}^{2} -2logL_i\right)$$

siendo i = 1,2, dónde 1 = seguro de hogar y 2 = seguro de automóvil:

"k_i": Número de parámetros de los modelos estimados,

"$logL_i$": Logaritmo neperiano del máximo valor de la función de verosimilitud para los modelos estimados.

3) Base de datos

La base de datos con la que se va a trabajar es una muestra que no contiene datos identificativos, para garantizar la anonimidad de los datos, de una cartera cedida a la Universidad de Barcelona por una gran entidad aseguradora multinacional.

Dicha muestra consta de información relativa a una cartera con dos productos; un seguro de hogar y un seguro de automóvil. Para una mejor comprensión de las variables que forman esta base de datos, se procede a explicar de qué tratan los seguros de hogar y los seguros de automóvil.

- Los seguros de hogar son un tipo de seguros que se fundamentan en la protección y cobertura de los posibles daños que se pudieran ocasionar en el hogar de los asegurados, ya sea en la vivienda propiamente dicha (continente) o en el mobiliario que hay dentro de la misma (contenido).
  Debido a esta distinción entre continente y contenido, se debe especificar en ambos casos el valor por el cual se asegurará cada uno, de manera que puede haber un límite de capital asegurado diferente para el contenido de la casa y el propio continente. Dentro del capital asegurado para el contenido irían el mobiliario, los electrodomésticos, la ropa e incluso los comestibles, teniendo en cuenta que joyas y otros objetos de lujo muchas veces deben especificarse aparte. Por otro lado, en el capital asegurado para el continente se incluyen las paredes, las puertas, y a menudo las ventanas y los vidrios, a veces teniendo que contratar una cobertura adicional para estos últimos.
  Los seguros de hogar son lo que se denomina como seguros multirriesgos, es decir, los riesgos que cubre dicho seguro pueden ser muy diversos; un incendio, un robo, la reparación de algún desperfecto que se pueda ocasionar en la casa, la responsabilidad civil que se deriva de la misma, o incluso la asistencia sanitaria que se derive de algún accidente ocurrido en el hogar.

- Los seguros de automóvil surgen de la necesidad de cubrir la responsabilidad civil derivada de los daños a terceros que se pueden producir cuando se conduce un vehículo a motor y, por tanto, para resarcir los posibles daños provocados por dicho conductor a terceros.
  Actualmente este seguro es obligatorio, por lo que todo propietario de un vehículo a motor debe contratarlo obligatoriamente, a no ser que sea contratado por otra persona que tenga interés en asegurar el vehículo.
  Además, es un seguro multirriesgo, ya que se pueden contratar coberturas voluntarias adicionales como pueden ser las coberturas por rotura de lunas, por incendio, por robo, o bien por daños propios.

Debido a las diferentes posibilidades de contratación y a la diversidad de riesgos que pueden ser cubiertos, se necesita obtener información de una serie de variables o factores de riesgo que permitan determinar las características de los riesgos asegurados y el cliente.

En la muestra de cartera de seguros con la que se va a trabajar disponemos de dieciocho variables, de la cuales dos son comunes para ambos seguros, ya que son información relativa al cliente, siete son variables pertenecientes al seguro de hogar y nueve son variables relacionadas con el seguro de automóviles. El número de registros que forman la muestra de la cartera es 10.041, teniendo una muestra suficientemente consistente como para obtener resultados sólidos. Además, de estas dieciocho variables, nueve son cualitativas y las otras nueve son cuantitativas.

Antes de proceder a analizar cada variable una por una, se muestra una tabla resumen con las variables que forman parte de la base de datos.

VARIABLES	BREVE DESCRIPCIÓN	TIPO VARIABLE	RELACIÓN
sexo	Sexo del cliente	Cualitativa	CLIENTE
edad	Edad del cliente	Cuantitativa	
estado_hogar	Estado póliza hogar	Cualitativa	HOGAR
tipo_hogar	Tipo de inmueble	Cualitativa	
cap_continente	Capital asegurado continente	Cuantitativa	
cap_contenido	Capital asegurado contenido	Cuantitativa	
mediador_hogar	Tipo de mediador pólizas hogar	Cualitativa	
forma_pago_hogar	forma de pago póliza hogar	Cualitativa	
antig_hogar	Antigüedad póliza hogar	Cuantitativa	
estado_auto	Estado póliza automóvil	Cualitativa	AUTO
tipo_auto	Tipo de vehículo	Cualitativa	
potencia_auto	Potencia vehículo	Cuantitativa	
edad_auto	Antigüedad vehículo	Cuantitativa	
asientos_auto	Número de asientos	Cuantitativa	
mediador_auto	Tipo de mediador pólizas automóvil	Cualitativa	
bonus_malus_auto	Bonificación/penalización póliza automóvil	Cuantitativa	
forma_pago_auto	Forma de pago poliza automóvil	Cualitativa	
antig_auto	Antigüedad póliza auto	Cuantitativa	

[Tabla 1] Tabla resumen variables cartera. Fuente: elaboración propia.

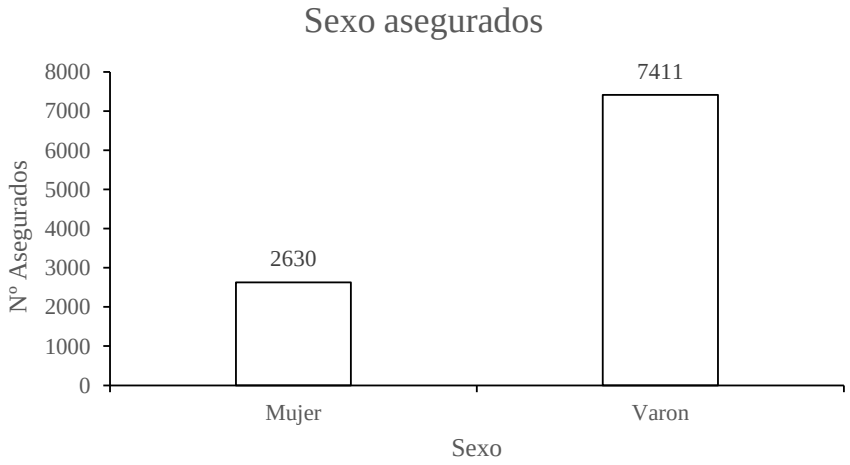
A continuación, se explica en qué consiste y que valores puede tomar cada variable junto con un gráfico que muestra la distribución de la misma.

3.1) Las variables de la cartera

Las primeras variables explicativas que se describen son las variables relacionadas con los clientes, las cuales son el sexo, que es una variable cualitativa, y la edad, la cual es una variable cuantitativa.

Debido a la entrada en vigor el 21 de diciembre de 2012 (tras el plazo de 5 años de moratoria) de la directiva del Consejo 2004/113/CE de 13 de diciembre de 2004 que prohíbe la distinción por sexos en el cálculo del precio de las primas, la variable **sexo** ya no se puede utilizar para las modelizaciones relacionadas con la tarificación de seguros. En este caso, al ser un estudio sin repercusión ninguna hacia los clientes, se usa dicha variable para ver si puede tener incidencia o no. Así pues, la variable sexo nos informa sobre el género del cliente, habiendo dos opciones posibles; mujer o varón.

La distribución del sexo es la siguiente:

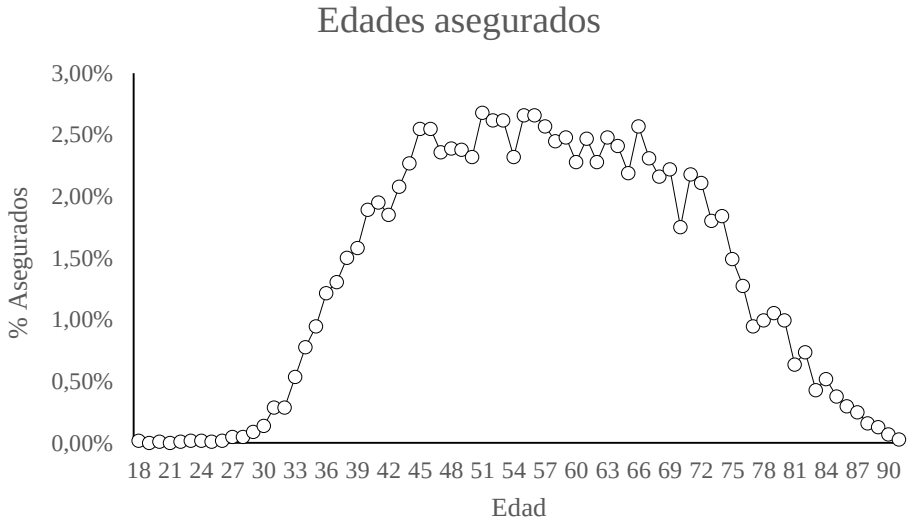


[Gráfico 1] Distribución del género de los asegurados. Fuente: elaboración propia.

Tal y como podemos ver en el [Gráfico 1], la mayoría de los asegurados son varones, ya que éstos representan casi el 74% de los asegurados en la cartera, mientras que las mujeres sólo representan el 26% restante.

Por otra parte, la **edad** es una variable importante, ya que dependiendo de la edad del cliente éste puede tener una mayor o menor propensión a la renovación de la póliza, en base a la hipótesis de que dependiendo de la edad cada variable tendrá un peso diferente sobre la decisión del cliente. Por ejemplo, los más jóvenes acostumbran a buscar el precio más competitivo, mientras que los clientes más adultos, acostumbran a valorar más que el producto en cuestión sea completo, es decir, buscan una mayor comodidad y protección, sin ser el precio el factor más decisivo en la mayoría de los casos.

La distribución de la variable edad es la siguiente:



[Gráfico 2] Distribución de las edades de los asegurados. Fuente: elaboración propia.

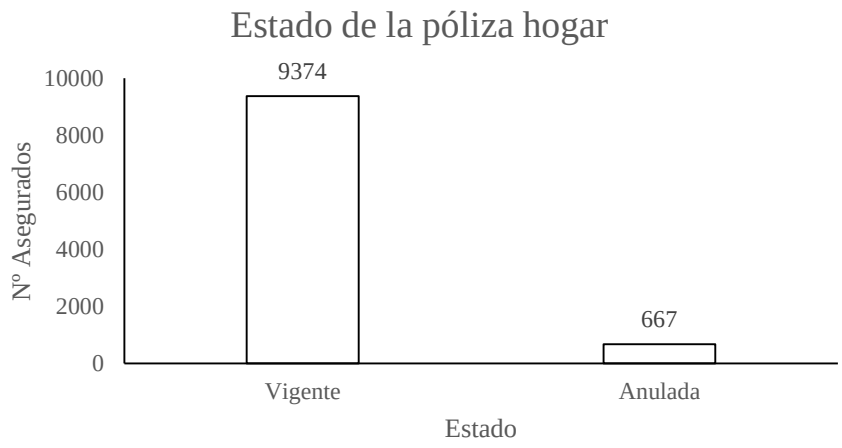
Como se puede observar en el [Gráfico 2], la variable edad tiene un rango de respuesta de entre dieciocho y noventa y un años.

Dentro de este rango, podemos afirmar que el intervalo de edades que más representación tienen en la cartera es el formado por los asegurados que tienen entre 38 y 75 años, los cuales representan más de un 85% del total de la cartera. Esto puede ser debido a que son las personas de entre estas edades las que acostumbran a tener las capacidades económicas suficientes - y psicomotrices en el caso de los automóviles - para tener un inmueble y un vehículo y, por tanto, para asegurarlos.

El siguiente grupo es el que contiene las variables relacionadas con el seguro de hogar. Las siete variables que lo componen son el estado de la póliza de hogar, el tipo de piso, el capital asegurado de continente, el capital asegurado de contenido, el tipo de mediador de las pólizas de hogar, la forma de pago de la póliza de hogar y la antigüedad de la póliza de hogar. Las dos variables de capital y la variable antigüedad son variables cuantitativas, mientras que el resto son cualitativas. La variable estado de la póliza de hogar es una de las dos variables dependientes de la cartera, siendo el resto variables explicativas.

El **estado de la póliza de hogar** es la variable que nos determina si los clientes tienen la póliza de hogar en vigor o si, por el contrario, la han anulado. Esta variable viene explicada por el resto de variables, las cuales son los factores que condicionan la decisión del cliente de seguir o no seguir en la compañía. Las dos opciones de respuesta posibles son “V”, si la póliza de hogar está vigente, o “A”, si la póliza está anulada.

La distribución de la variable estado de la póliza de hogar es la siguiente:

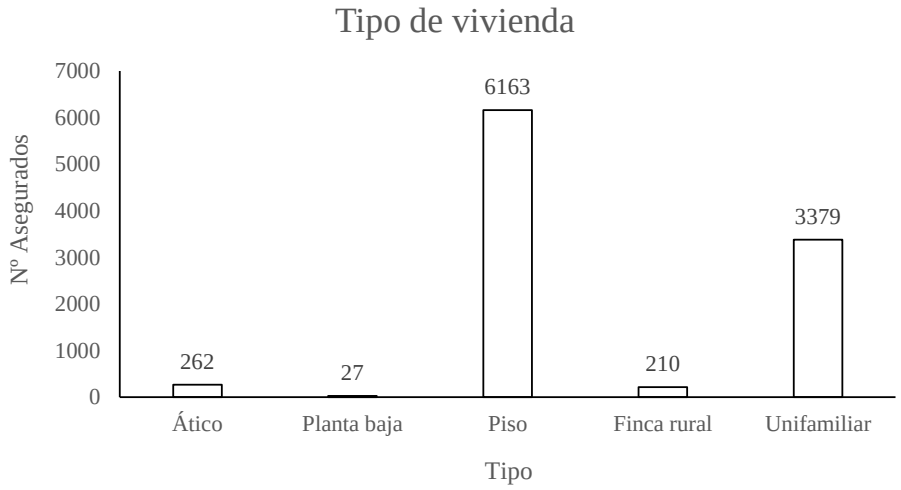


[Gráfico 3] Distribución del estado de las pólizas de hogar. Fuente: elaboración propia.

En este [Gráfico 3] podemos apreciar como la mayoría de las pólizas de hogar presentes en la base de datos están vigentes, estando anuladas sólo un 6,65% y, por tanto, representando las pólizas vigentes un 93,4%.

El **tipo de piso** es la variable que nos indica de qué clase de vivienda se ha asegurado el riesgo, es decir, qué características tiene el objeto asegurado. El riesgo asociado al inmueble puede ser de diferente naturaleza dependiendo si es un piso en una ciudad, o si es una vivienda unifamiliar. Las posibles respuestas de esta variable son, "PI", cuando se trata de un piso, "UF" y "UA", cuando es una casa unifamiliar, "PB", cuando es una planta baja, "AT", cuando la vivienda es un ático y "RU", cuando es una finca rural.

La distribución de la variable tipo de piso es la siguiente:

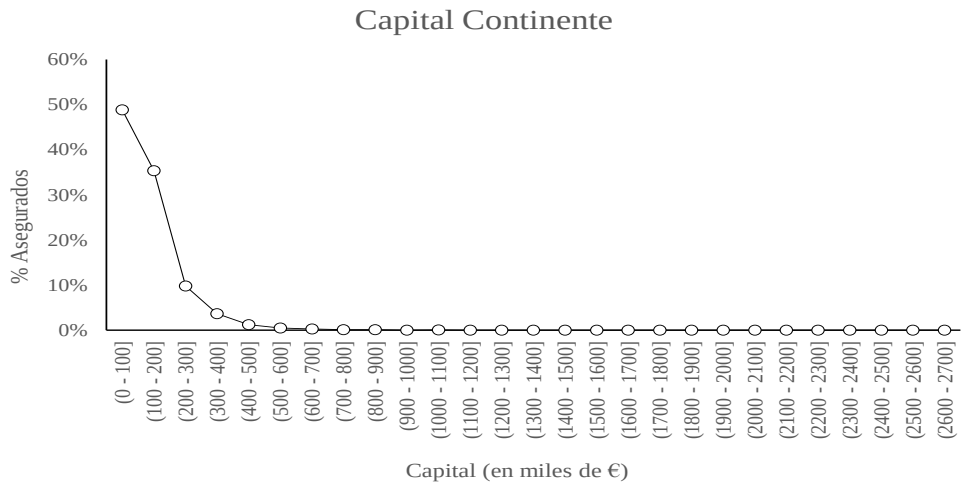


[Gráfico 4] Distribución del tipo de vivienda. Fuente: elaboración propia.

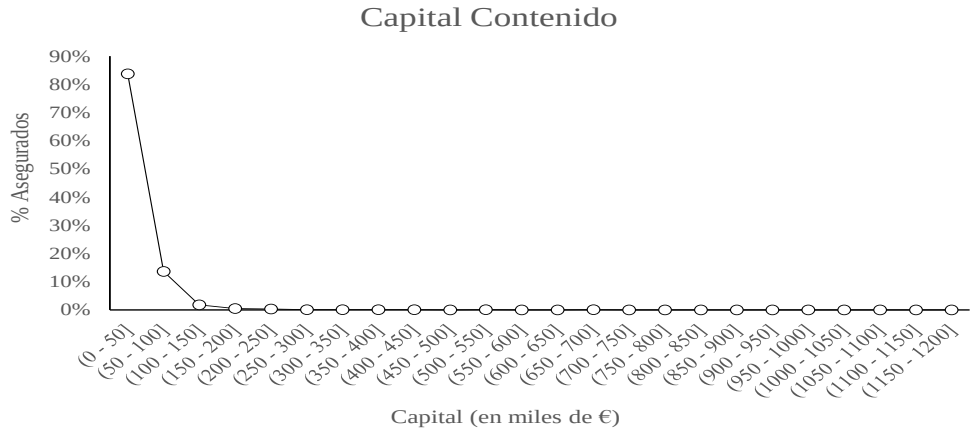
Tal y como podemos observar en el [Gráfico 4], el tipo de vivienda más habitual es el piso, con un 61,4% de la cartera, seguido por la vivienda unifamiliar, con un 33,7% del total. Por último, tanto los áticos, las fincas rurales y las plantas bajas son un tipo de viviendas poco comunes, representando menos de un 5% todas juntas.

Las variables **capital asegurado continente y contenido** determinan qué cantidad del inmueble, expresada en cuantía monetaria, queda cubierta por el seguro. En el caso de la variable cuantía de continente asegurado, se especifica el valor máximo del contenedor (si entendemos el inmueble como un recipiente) que se pagará si ocurre algún siniestro contemplado en las coberturas de la póliza contratada. Por otro lado, la variable cuantía del contenido asegurado nos expresa el valor máximo de lo que se encuentra dentro del contenedor que será repuesto por la aseguradora en caso de ocurrencia de siniestro.

Dichas variables son interesantes para ver qué incidencia puede tener en la retención del cliente que éste asegure una cuantía mayor o menor. La distribución de las variables capital asegurado continente y contenido son las siguientes:



[Gráfico 5] Distribución del capital continente. Fuente: elaboración propia.



[Gráfico 6] Distribución del capital contenido. Fuente: elaboración propia.

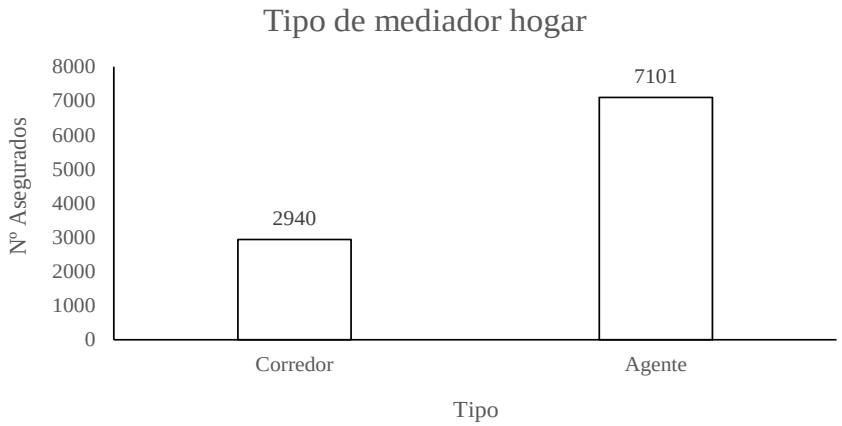
En el [Gráfico 5] podemos observar como el capital por continente asegurado, casi en el 50% de los casos, no llega a los cien mil €. Si tenemos en cuenta que el valor de la mayoría de las viviendas normales no sobrepasa los quinientos mil €, es lógico que las pólizas con un capital asegurado por continente menor a quinientos mil € represente un 98,84% del total de la cartera.

Por otro lado, en el [Gráfico 6] podemos apreciar como alrededor del 84% de los clientes han asegurado un capital por contenido inferior a cincuenta mil €, mientras que solo un 13,6% ha asegurado entre cincuenta y cien mil €, siendo casi residual el porcentaje de asegurados que han contratado una cobertura de capital contenido superior a cien mil €.

Al tratarse de una compañía mediada, la variable **tipo de mediador hogar** nos especifica qué tipo de mediador ha logrado realizar la colocación del producto asegurador, habiendo dos posibilidades; que sea una agente (normalmente exclusivo), o bien que sea un corredor de seguros.

Tanto para los seguros de hogar como para los seguros de automóvil (que veremos a continuación), los agentes de seguros acostumbran a ser trabajadores que trabajan por cuenta de la entidad aseguradora y que, por tanto, sólo ofrecen a sus clientes los productos de la entidad por la que trabajan. Por otra parte, los corredores de seguros acostumbran a trabajar para diferentes compañías aseguradoras y, por tanto, ofrecen un abanico más amplio de posibilidades de contratación a sus clientes. Es muy interesante poder ver cómo éstos pueden interferir o decantar la decisión del consumidor hacia un lado u otro.

La distribución de la variable tipo de mediador hogar es la siguiente:

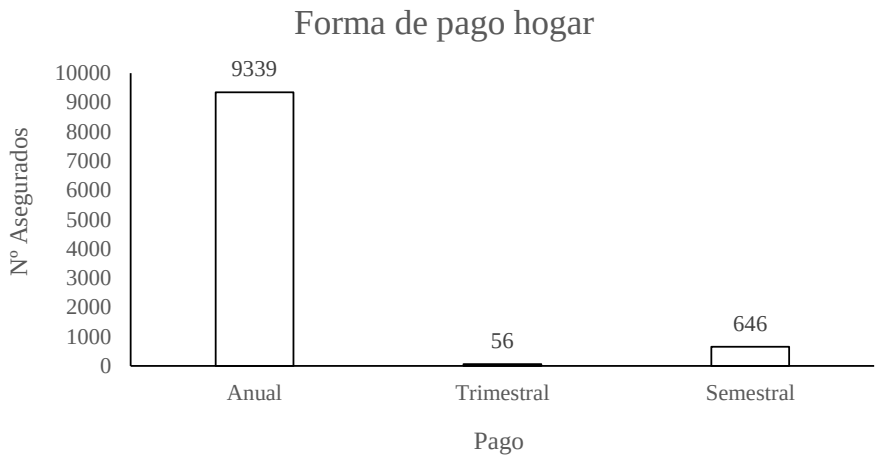


[Gráfico 7] Distribución del tipo de mediador de la póliza de hogar. Fuente: elaboración propia.

El [Gráfico 7] nos muestra la distribución del tipo de mediador de la póliza de hogar, viéndose claramente como el 70% de las pólizas han sido realizadas por los agentes de seguros, mientras que 30% restante ha sido colocado por los corredores de seguros - los cuales, a diferencia de los agentes, tienen un amplio abanico de posibilidades para comercializar -.

La **forma de pago de la póliza de hogar** es una variable que nos indica de que manera ha pagado el cliente. Tanto para los seguros de hogar como los de automóvil (que veremos a continuación), es una variable que puede ser interesante para ver que incidencia puede tener en la decisión de cada tipo de cliente tener una forma de pago diferente y, por tanto, hasta que punto influye que paguen de una sola vez o bien que paguen con una cierta temporalidad. Las posibles opciones de respuesta de esta variable son; "A" si es anual, "S" si es semestral o "T" si es trimestral.

La distribución de la variable forma de pago de la póliza de hogares es la siguiente:

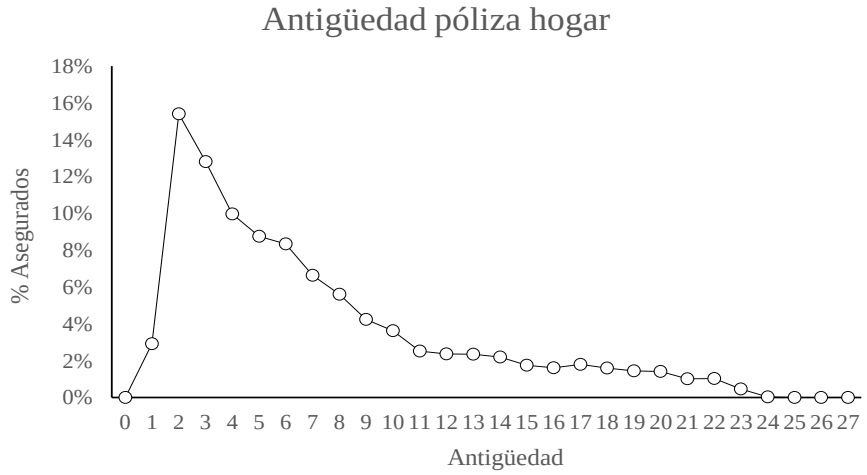


[Gráfico 8] Distribución de la forma de pago de la póliza de hogar. Fuente: elaboración propia.

La forma de pago más común, tal como nos indica el [Gráfico 8], es el pago anual, representando el 93% de la cartera, mientras que el pago semestral es usado por el 6,5% de los asegurados aproximadamente y el pago trimestral apenas es utilizado por el 0,5% restante.

Por último, la variable **antigüedad de la póliza de hogar** informa sobre la cantidad de años que lleva en vigor cada póliza. Es un factor determinante, ya que en general, los clientes con pólizas más antiguas tienden a anular menos, ya que llevan mucho tiempo en la compañía, lo que supone que están cómodos y contentos con el servicio prestado y no son propensos a cambiar de compañía con facilidad.

La distribución de la variable antigüedad de la póliza de hogar es la siguiente:



[Gráfico 9] Distribución de la antigüedad de la póliza de hogar. Fuente: elaboración propia.

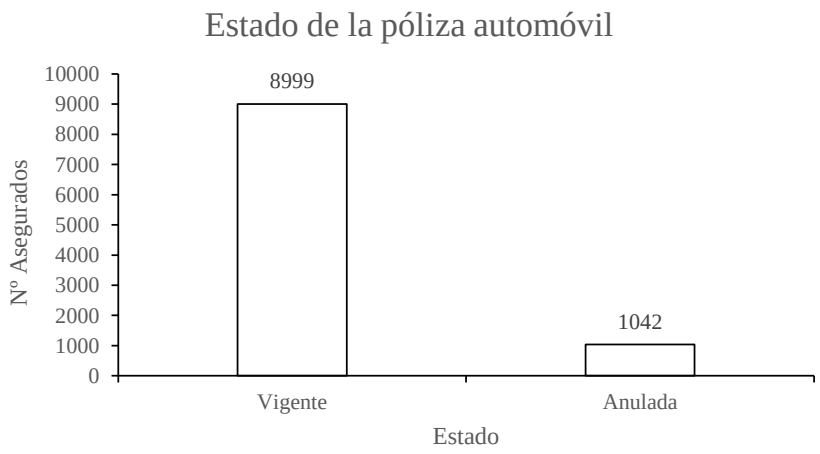
Si se observa el [Gráfico 9], se puede comprobar como la variable antigüedad de la póliza de hogar tiene un rango de respuesta de entre uno y veintisiete años. La mayoría de los asegurados, los cuales representan alrededor del 78,4% de la cartera, hace entre uno y diez años que contrataron la póliza, mientras que los que llevan más de diez años en la compañía representan el 21,6% restante. Si tenemos en cuenta que los inmuebles tienen una vida útil más larga que los vehículos, es lógico ver como la duración media de estas pólizas es mayor que las de los automóviles, tal y como veremos posteriormente.

El tercer y último grupo es el que contiene las variables relacionadas con el seguro de automóvil. Las nueve variables que lo componen son el estado de la póliza de automóvil, el tipo de vehículo, la potencia del vehículo, la antigüedad del vehículo, el número de asientos del vehículo, el tipo de mediador de las pólizas de automóvil, la bonificación/penalización de la póliza de automóvil, la forma de pago de la póliza de automóvil y por último la antigüedad de la póliza de automóvil.

Las variables potencia, antigüedad del vehículo, número de asientos y antigüedad de la póliza de automóvil son variables cuantitativas, mientras que el resto son cualitativas. La variable estado de la póliza de automóvil es una de las dos variables dependientes de la cartera, siendo el resto variables explicativas.

El **estado de la póliza de automóvil** es la variable que nos determina si los clientes tienen la póliza de automóvil en vigor, o si por el contrario, la han anulado. Esta variable, como ocurría con la variable estados de la póliza de hogar, viene explicada por el resto de variables, las cuales son los factores que condicionan la decisión del cliente de seguir o no seguir en la compañía. Las dos opciones de respuesta posibles son; “V”, si la póliza de automóvil está vigente, o “A”, si la póliza está anulada.

La distribución de la variable estado de la póliza de automóvil es la siguiente:

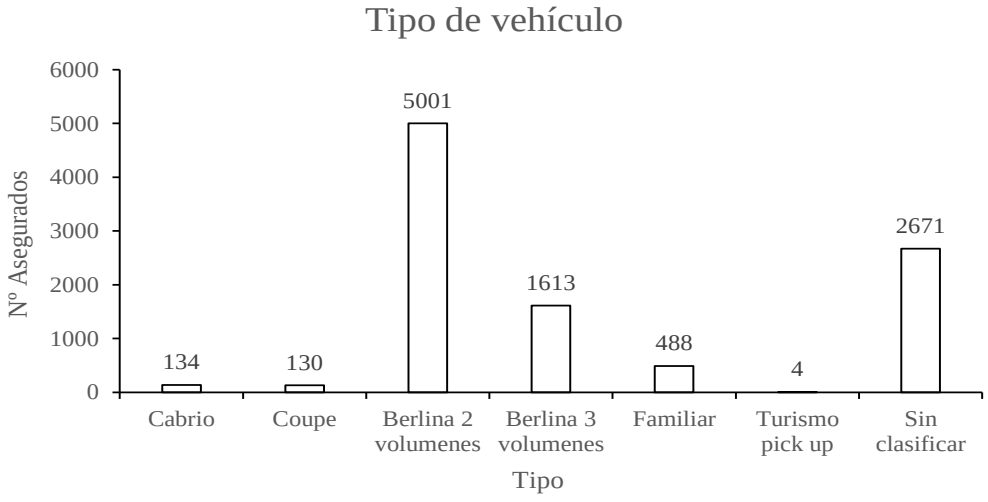


[Gráfico 10] Distribución del estado de la póliza de automóvil. Fuente: elaboración propia.

En este [Gráfico 10] podemos observar como el porcentaje de pólizas vigentes es mucho superior que el de pólizas anuladas, tal y como sucedía con las pólizas de hogar, aunque en este caso con un porcentaje ligeramente inferior, situándose casi en el 90% del total, y por tanto, representando las pólizas anuladas el 10% restante.

El **tipo de vehículo** nos informa qué clase de vehículo es el que se asegura en cada póliza. Esta variable nos puede proporcionar información muy interesante, ya que dependiendo del tipo de vehículo que tenga el asegurado se puede dilucidar qué tipo de uso se le da al vehículo, pudiendo ser, en términos generales, un uso más recreativo y de ocio cuando es un vehículo "cabrio" o "coupé", un uso más familiar cuando es un vehículo "berlina" o "familiar", o bien un uso de trabajo cando es un vehículo "turismo pick up". Así pues, las posibles opciones de respuesta de esta variable son; "berlina 2 volúmenes", "berlina 3 volúmenes", "cabrio", "coupé", "familiar", "turismo pick up" o "NA" cuando no está informado.

La distribución de la variable tipo de vehículo es la siguiente:



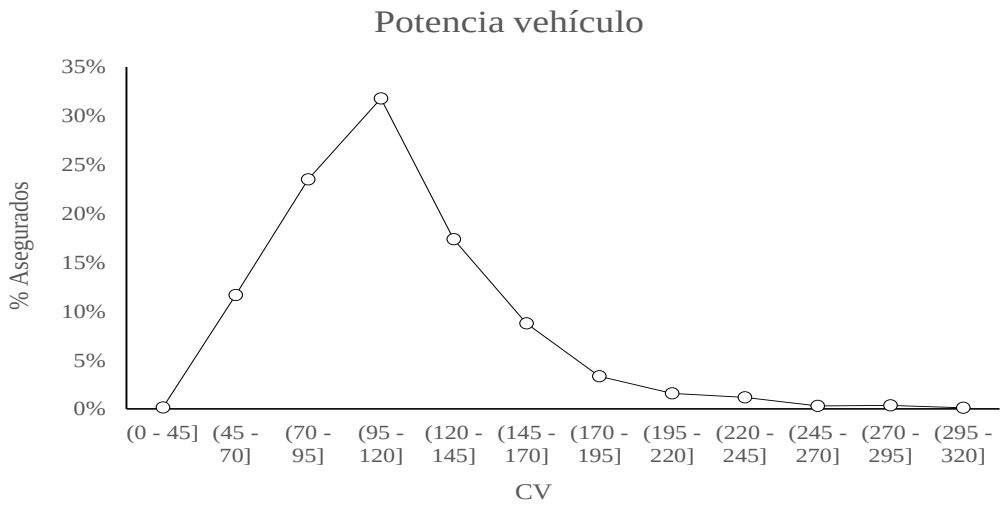
[Gráfico 11] Distribución del tipo de vehículo. Fuente: elaboración propia.

En este [Gráfico 11] podemos observar como las berlinas de 2 y 3 volúmenes son el tipo de coche favorito de los asegurados, representando entre ambas casi un 66% del total de vehículos. El resto de tipos de vehículo aglutinan un 7,5% del total de vehículos, mientras que el 26,5% restante corresponde a vehículos sin clasificar.

La siguiente variable es la **potencia del vehículo**, la cual nos indica los caballos de vapor (CV) del vehículo, unidad de medida de potencia empleada en el sector automovilístico.

Esta variable podría ser de utilidad, ya que dependiendo de la potencia del vehículo, se puede tener una idea aproximada del estilo de conducción del conductor, ya que por lo general, aquel conductor que prefiere un vehículo con una potencia elevada, o bien será un conductor más agresivo y que disfruta de la velocidad, o bien necesita dicha potencia para llevar un vehículo de dimensiones y peso importantes.

La distribución de la variable potencia del vehículo es la siguiente:



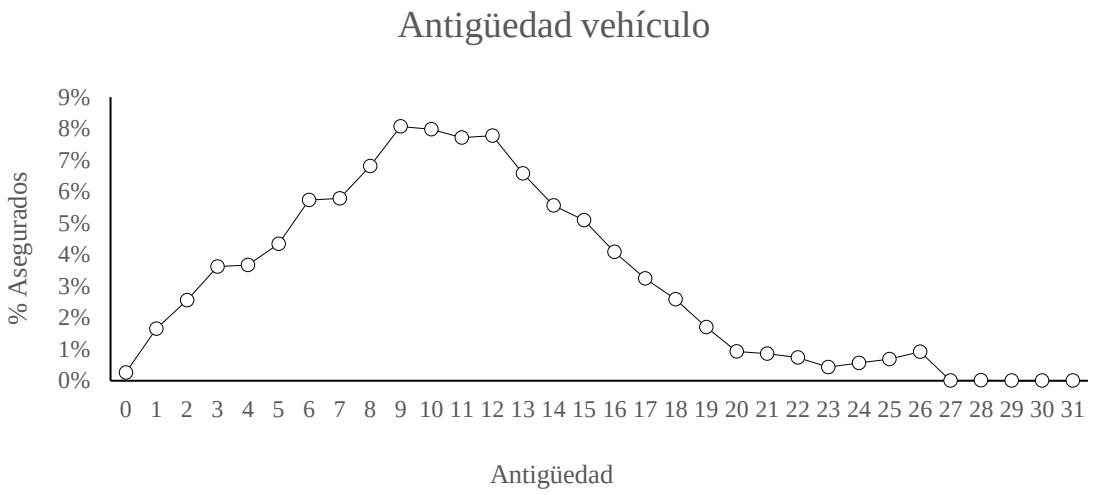
[Gráfico 12] Distribución de la potencia del vehículo. Fuente: elaboración propia.

Como se puede observar en el [Gráfico 12], la variable potencia del vehículo tiene un rango de respuesta de entre cuarenta y cinco y trescientos caballos de vapor. Los intervalos con más peso en la cartera son lo que van entre los setenta y los ciento setenta caballos de vapor, lo cual tiene cierta lógica si se tiene en cuenta que se corresponde con las potencias más comunes en el mercado, representando el 81.4% de todos los vehículos asegurados. Los vehículos con una potencia considerable están situados por encima de los ciento setenta caballos de vapor, lo que equivale al 6,9% del total, siendo los vehículos con menos de setenta caballos de vapor el 11.7% restante.

La variable **antigüedad del vehículo** determina el número de años que han transcurrido des de que el vehículo se matriculó y, por tanto, cuantos años hace que se está utilizando.

Aunque el vehículo puede haber tenido temporadas de poco uso o casi inexistente, en términos generales, cuanto mayor es su antigüedad menos seguro es, bien sea por el desgaste que han sufrido sus componentes a lo largo del tiempo, o bien sea porque no incorpora las últimas novedades en materia de seguridad y, por tanto, acostumbra a tener una mayor probabilidad de siniestro.

La distribución de la variable antigüedad del vehículo es la siguiente:

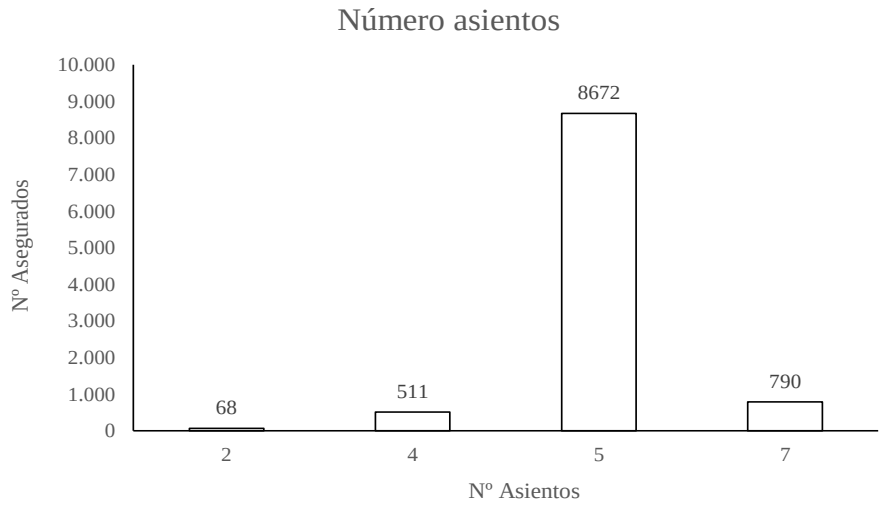


[Gráfico 13] Distribución de la antigüedad del vehículo. Fuente: elaboración propia.

La variable antigüedad del vehículo tiene un rango de respuesta de entre cero (es decir, el vehículo es nuevo) y treinta y un años, tal y como se muestra en el [Gráfico 13]. Los vehículos de entre ocho y trece años de antigüedad son los más comunes, representando el 45% del total de vehículos, lo cual se corresponde con la media de años del parque móvil español. El otro 55% se divide entre los vehículos más nuevos - que tienen menos de 8 años -, los cuales representan casi un 28% del total, y entre los vehículos más antiguos - que tienen más de 13 años -, los cuales suponen el 27% restante.

La siguiente variable es el **número de asientos del vehículo**, la cual nos da información acerca de las plazas del automóvil y de cuanta gente puede llegar a albergar. Esta variable nos puede dar información de si el cliente busca un vehículo amplio, y con una buena capacidad para transportar muchas personas, o si bien prefiere un vehículo con menos capacidad y quizás más deportivo. Las posibles opciones de respuesta de esta variable son: dos, cuatro, cinco o siete asientos.

La distribución de la variable número de asientos del vehículo es la siguiente:

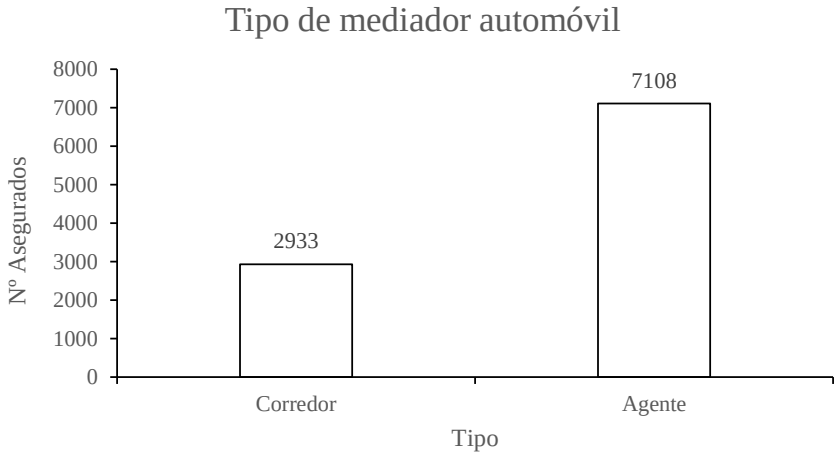


[Gráfico 14] Distribución del número de asientos del vehículo. Fuente: elaboración propia.

La variable número de asientos tiene una distribución muy desigual, ya que en el [Gráfico 14] se puede observar que la gran mayoría de los vehículos tienen cinco asientos, suponiendo el 86,4% del total de vehículos. Le siguen los vehículos grandes que pueden tener siete asientos con un 7,9%, a continuación los vehículos con cuatro asientos que representan el 5,1% de los vehículos asegurados y, por último, los que tienen dos asientos, con un ínfimo 0,6%.

Tal y como se ha comentado en el caso de los seguros de hogar, al tratarse de una compañía mediada, la variable tipo de mediador automóvil nos especifica qué tipo de mediador ha logrado realizar la colocación del producto asegurador, habiendo dos posibilidades: que sea una agente (normalmente exclusivo), o bien que sea un corredor de seguros.

La distribución de la variable tipo de mediador automóvil es la siguiente:

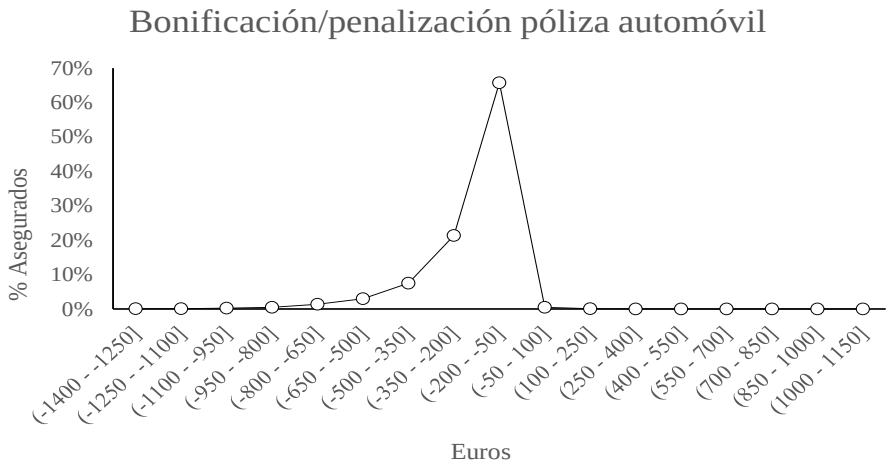


[Gráfico 15] Distribución del tipo de mediador de las pólizas de automóvil. Fuente: elaboración propia.

Según lo dispuesto en el [Gráfico 15], la distribución de la variable tipo de mediador para las pólizas de automóvil es casi idéntica a la distribución de las pólizas de hogar. En este caso, las pólizas comercializadas por los agentes representan el 71% del total de pólizas, mientras que el 29% restante corresponde a las vendidas por corredores.

La variable **bonificación/penalización de la póliza de automóvil** nos indica la cantidad monetaria que ha sido bonificada o penalizada sobre la prima inicial, teniendo en cuenta que cuando el valor está en negativo es un "bonus" y cuando está en positivo significa "malus". Esta variable nos aporta una información fundamental, ya que si al tomador de la póliza le aplican penalizaciones posiblemente sea más propenso a no anular el seguro debido a su alta siniestralidad, mientras que si le aplican pocas bonificaciones puede que cambie a una compañía aún más barata. Además, dependiendo de la cuantía de cada "bonus/malus", la probabilidad de anulación será diferente en cada caso.

La distribución de la variable bonificación/penalización de la póliza de automóvil es la siguiente:

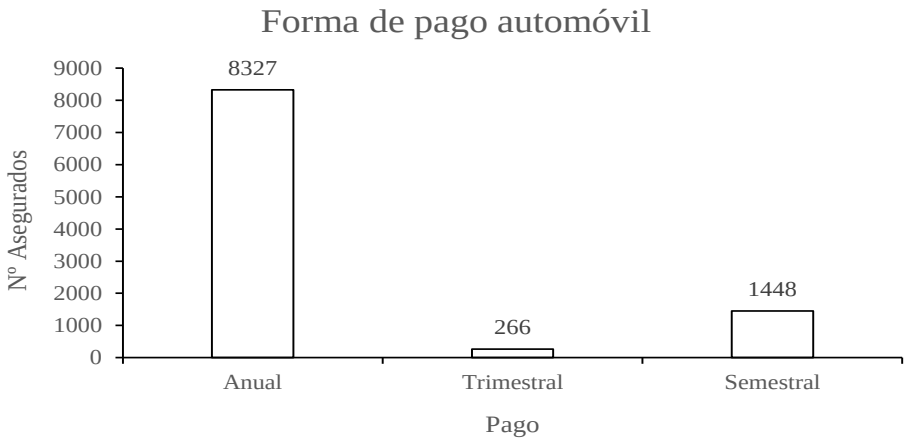


[Gráfico 16] Distribución de los "Bonus/Malus" de las pólizas de automóvil. Fuente: elaboración propia.

La variable bonificación/penalización de la póliza de automóvil tiene un rango de respuesta de entre menos mil trescientos noventa y seis con diez y mil veinte con sesenta y dos euros. Como se puede observar en el [Gráfico 16], a la mayoría de los asegurados, los cuales representan el 87% del total, se les ha aplicado un "bonus" de entre cincuenta y trescientos cincuenta €, lo que quiere decir que posiblemente no hayan tenido mucha siniestralidad o casi nula. Por otro lado, los que han tenido un gran "bonus" suponen el 12,5% de la cartera, mientras que aquellos que han sido penalizados o se les ha aplicado "malus" son el 0.5% restante.

La penúltima variable relacionada con el seguro de automóvil es la **forma de pago de la póliza de automóvil**, que nos indica, como en el caso de los seguros de hogar, con que temporalidad ha pagado el cliente la prima de la póliza.

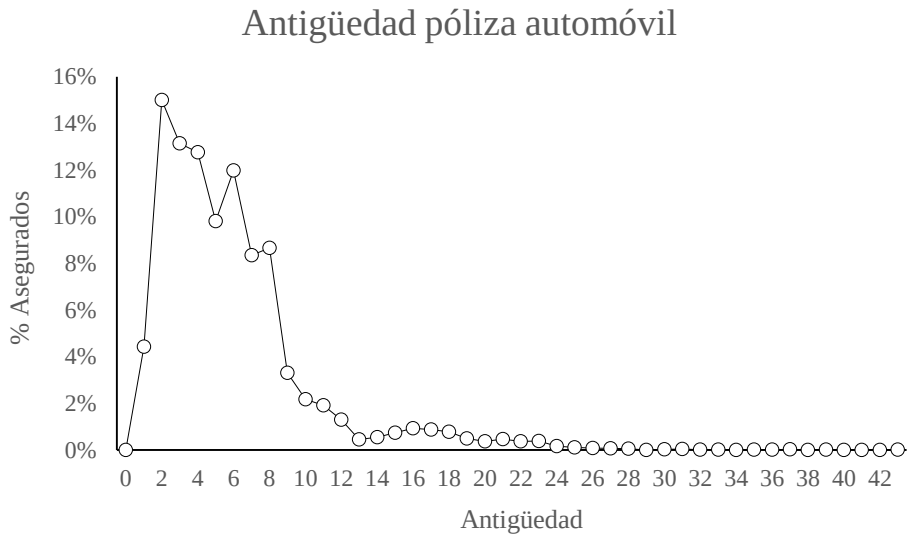
Las posibles opciones de respuesta de esta variable son; “A” si es anual, “S” si es semestral, o “T” si es trimestral. La distribución de la variable forma de pago de las pólizas de automóvil es la siguiente:



[Gráfico 17] Distribución de la forma de pago de las pólizas de automóvil. Fuente: elaboración propia.

Tal y como ocurría con las pólizas de hogar y como se puede observar en el [Gráfico 17], la forma de pago más común es el pago anual, aunque en este caso representando el 83% de la cartera, mientras que el pago semestral es usado por el 14,4% de los asegurados aproximadamente y el pago trimestral por el 2,6% restante

Por último, la variable **antigüedad de la póliza de automóvil** informa sobre la cantidad de años que lleva en vigor cada póliza de automóvil. La antigüedad de la póliza no tiene por que ser obligatoriamente igual o menor a la antigüedad del vehículo, ya que con la misma póliza se pueden haber ido asegurando diferentes vehículos a lo largo de los años. Evidentemente, los datos de los vehículos con los que se trabajan son los del vehículo actual. La distribución de la variable antigüedad de la póliza de automóvil es la siguiente:



[Gráfico 18] Distribución de la antigüedad de las pólizas de automóvil. Fuente: elaboración propia.

En este caso, tal y como está graficado en el [Gráfico 18], la variable antigüedad de la póliza de automóvil tiene un rango de respuesta de entre uno y cuarenta y tres años. El rango de antigüedad más común es el que abarca las pólizas de cero a diez años, el cual representa el 89.6% del total de las pólizas, mientras que el resto de pólizas - que por tanto tienen más de 10 años de antigüedad – suponen solamente el 10,4% restante. Así pues, en comparación con las pólizas de hogar, las de automóvil tienen una mayor concentración en el primer tramo de cero a diez años y se corrobora que debido a que el mercado de los seguros de automóvil tiene más movilidad, no acostumbran a tener una duración tan larga.

4) Aplicación práctica; modelización y análisis de los resultados

Una vez explicados los modelos que vamos a usar y conocidas las características de la cartera a modelizar, es el momento de aplicar estos conocimientos para modelizar la retención de los clientes. Algunas respuestas de las variables cualitativas son similares y tienen poca frecuencia, por lo que para proceder a la modelización de dichas variables es preferible dicotomizarlas, lo que supone agrupar sus posibles respuestas por características en dos grupos diferentes. En la siguiente tabla se puede ver la dicotomización efectuada, teniendo en cuenta que si existen respuestas "desconocido" serán añadidas a la categoría más frecuente de la variable.

	Variables Cualitativas		
	Respuesta Original	**Nueva respuesta**	**Significado**
Sexo	Mujer	0	Mujer
	Varón	1	Hombre
Estado Hogar	"A" Anulada	0	Anula
	"V" Vigente	1	Renueva
Tipo Hogar	"RU" Rural "UF" Unifamiliar	0	Vivienda Unifamiliar
	"PB" Planta baja "AT" Ático "PI" Piso	1	Piso
Mediador Hogar	Corredor	0	Corredor
	Agente	1	Agente
Forma de Pago Hogar	"T" Trimestral "S" Semestral	0	Fraccionado
	"A" Anual	1	Anual
Estado Auto	"A" Anulada	0	Anula
	"V" Vigente	1	Renueva
Tipo Auto	Cabrio Coupé Turismo Pick-Up	0	Uso diferente a familiar
	Berlina 2 Volúmenes Berlina 3 Volúmenes Familiar	1	Uso Familiar
Mediador Auto	Corredor	0	Corredor
	Agente	1	Agente
Forma de Pago Auto	"T" Trimestral "S" Semestral	0	Fraccionado
	"A" Anual	1	Anual

[Tabla 2] Cuadro Resumen de la transformación de las variables cualitativas. Fuente: elaboración propia.

Para poder hacerse una idea más concreta de los datos que finalmente se van a modelizar, a continuación se muestran dos tablas con los estadísticos descriptivos de las variables cualitativas y cuantitativas que han sido tratadas y que, por tanto, van a ser objeto de la modelización.

	Descriptivos Variables Cualitativas Dicotomizadas			
	Respuesta	**Significado**	**Frecuencia**	**Porcentaje**
Sexo	0	Mujer	2630	26,19
	1	Hombre	7411	73,81
Estado Hogar	0	Anula	667	6,64
	1	Renueva	9374	93,36
Tipo Hogar	0	Vivienda Unifamiliar	3589	35,74
	1	Piso	6452	64,26
Mediador Hogar	0	Corredor	2940	29,28
	1	Agente	7101	70,72
Forma de Pago Hogar	0	Fraccionado	702	6,99
	1	Anual	9339	93,01
Estado Auto	0	Anula	1042	10,38
	1	Renueva	8999	89,62
Tipo Auto	0	Uso diferente a familiar	2939	29,27
	1	Uso Familiar	7102	70,73
Mediador Auto	0	Corredor	2933	29,21
	1	Agente	7108	70,79
Forma de Pago Auto	0	Fraccionado	1714	17,07
	1	Anual	8327	82,93

[Tabla 3] Distribución de frecuencias de las variables cualitativas dicotomizadas. Fuente: elaboración propia.

Variable	Media	Desviación Estándar	Mínimo	Máximo	Moda	Suma
Edad	57,30	13,06	18	91	51	575.370
Capital Continente (*)	109,57	111,59	0,01	2.478,52	106,10	1.100.230,26
Capital Contenido (*)	31,03	33,22	0,0005	1.169,39	31,83	311.595,03
Antigüedad Hogar	7,13	5,29	1	27	2	71.560
Potencia Auto	1,13	0,38	0,45	3	0,90	11.351,51
Edad Auto	10,72	5,18	0	31	9	107.619
Asiento Auto	5,09	0,65	2	7	5	51.070
Bonus/Malus Auto	-205,48	144,06	-1.396,10	1.020,62	0	-2.063.248,88
Antigüedad Auto	5,96	4,51	1	43	2	59.806

(*) en miles de euros

[Tabla 4] Estadísticos descriptivos de las variables cuantitativas. Fuente: elaboración propia.

4.1) Tabla de contingencias

Una vez ya tenemos la base de datos que vamos a modelizar lista, y antes de proceder con la modelización de los modelos univariantes, cabe hacer un análisis de las interdependencias presentes entre las dos variables dependientes explicadas anteriormente. Así, si la relación entre las dos variables dependientes, sin tener en cuenta las variables explicativas, muestra dependencia, justificaría el uso de un modelo bivariante.

Una tabla de contingencias es una herramienta que nos permite detectar y medir las relaciones presentes entre diferentes variables de una manera simple y muy visual.

El funcionamiento de la tabla es muy fácil, ya que se basa en el cruce de unas variables "fila" con las variables "columna", donde cada intersección nos muestra la intensidad de la interrelación entre cada par de variables.

Tabla de Contingencias		Estado Póliza Automóvil		
		No Renueva	**Renueva**	**Total**
Estado Póliza Hogar	**No Renueva**	275	392	667
		2.74	3.90	6.64
		41.23	58.77	
		26.39	4.36	
	Renueva	767	8607	9374
		7.64	85.72	93.36
		8.18	91.82	
		73.61	95.64	
	Total	1042	8999	10041
		10.38	89.62	100.00

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 5] Tabla de contingencias de las variables explicadas. Fuente: elaboración propia.

Acorde con lo mostrado en la [Tabla 5], podemos observar como existe una gran incidencia en un seguro cuando el otro ha sido renovado anteriormente. Así, cuando el tomador renueva el seguro de hogar o de auto, en un 85,72% también renueva el otro seguro. Podemos ver también como el porcentaje de renovaciones del seguro de hogar asciende a un total del 93,36%, mientras que el seguro de auto solo es renovado en un 89,62% de los casos. Por otro lado, sólo en el 2,74% de los casos los tomadores no renuevan ninguno de los dos seguros.

Estadístico	DF	Valor	Prob.
Chi-Square	1	7.311.973	<.0001

[Tabla 6] Estadístico Chi-Square. Fuente: elaboración propia.

Además, si se precisa llevar a cabo un contraste de si existe dependencia entre las dos variables explicadas, el estadístico "Chi-Square" nos puede ser de gran ayuda. Como podemos ver en la [Tabla 6], el p-value es inferior a 0,05, por lo que con un nivel de confianza del 95% se rechaza la hipótesis nula de independencia y se pude afirmar que existe dependencia significativa entre las dos variables dependientes; estado de la póliza de hogar y estado de la póliza de auto.

De todas maneras, aunque exista dependencia entre las variables explicadas, es interesante realizar las modelizaciones univariantes para analizar cómo influyen las variables explicativas en las decisiones de los tomadores de renovar su póliza de hogar o de automóvil de manera independiente.

4.2) Modelización univariante

En este apartado se estiman los dos modelos univariantes, que son, como ya se ha comentado en el apartado 2.1, el modelo probit univariante y el modelo logit univariante.

Al ser modelos univariantes se modelizan los dos seguros por separado, es decir, por un lado se aplican los modelos logit y probit univariante a la subcartera formada por la variable explicada del seguro de hogar con sus variables explicativas (las cuales incluyen las variables comunes sexo y edad del tomador) y por otro lado se aplican a la subcartera formada por la variable explicada del seguro de automóvil con sus variables explicativas (incluyendo también las dos variables comunes mencionadas anteriormente).

En la siguiente [Tabla 7] podemos observar los diferentes niveles de significación con el que se contrastan los parámetros para saber si estos son significativos o no.

Al lado derecho de cada estimador de la [Tabla 8] está indicado el grado de significación de cada parámetro, pudiendo ser significativo a un nivel de significación del 10%, del 5% o del 1%.

Nivel de Significación	
***	1%
**	5%
*	10%

[Tabla 7] Niveles de significación contrastados. Fuente: elaboración propia.

Variables Explicativas / Modelos		Hogar		Automóvil	
		Probit	Logit	Probit	Logit
Comunes	Intercept	0,9779 ***	1,5616 ***	0,9565 ***	1,5843 ***
	Sexo	-0,0192	-0,0400	-0,1040 **	-0,2070 ***
	Edad	0,0020	0,0038	0,0039 ***	0,0074 ***
Hogar	Tipo Hogar	-0,0699	-0,1415		
	Capital Continente	-0,0001	-0,0003		
	Capital Contenido	0,0015 *	0,0037 **		
	Mediador Hogar	0,1317 ***	0,2662 ***		
	Forma de Pago Hogar	0,2815 ***	0,5683 ***		
	Antigüedad Hogar	0,0141 ***	0,0308 ***		
Auto	Tipo Auto			-0,0687 *	-0,1332 *
	Potencia Auto			0,0777	0,1489
	Edad Auto			-0,0319 ***	-0,0602 ***
	Asiento Auto			0,0573 **	0,1086 **
	Mediador Auto			0,2208 ***	0,4173 ***
	Bonus/Malus Auto			0,0006 ***	0,0011 ***
	Forma de Pago Auto			0,0622	0,1186
	Antigüedad Auto			0,0182 ***	0,0364 ***

[Tabla 8] Estimadores y contrastes de significación de los modelos univariantes. Fuente: elaboración propia.

Esta [Tabla 8] recoge los estimadores y los contrastes de significación de los parámetros asociados a las variables incluidas en los modelos logit y probit univariantes con los que se ha modelizado las subcarteras de los seguros de hogar y de automóvil respectivamente.

En los dos siguientes apartados se procede a analizar el signo de cada estimador y su contraste de significación.

4.2.1) Seguro de Hogar

Respecto al seguro de hogar, podemos observar en la [Tabla 8] como cuenta con nueve parámetros, de los cuales cuatro no son significativos en ninguno de los tres niveles de significación recogidos en la [Tabla 7] para ninguno de los dos modelos.

Si observamos el modelo probit univariante, el parámetro asociado a la variable capital asegurado contenido es significativo al 10%, pero no al 5% ni al 1%, mientras que, por otra parte, las variables mediador del seguro de hogar, forma de pago del seguro de hogar, antigüedad del seguro de hogar y el término constante son significativos en los tres niveles de significación contrastados.

En cuanto al signo de los estimadores de los parámetros que son significativos, podemos ver como para los dos parámetros asociados a las variables cuantitativas que son significativos en algún nivel de significación de los tres antes comentados, la relación es directa. Es decir, cuando aumenta el capital contenido asegurado o cuando mayor es la antigüedad de la póliza de hogar la probabilidad de renovación aumenta.

Alternativamente, en cuanto a las variables cualitativas se observa que cuando el mediador es un agente o cuando la forma de pago es anual las probabilidades de renovación también aumentan.

Por otro lado, si observamos el modelo logit univariante, llegamos a las mismas conclusiones que en el modelo probit univariante en lo que a significación de los parámetros estimados se refiere, con la única diferencia que en el caso que nos concierne el parámetro asociado a la variable sexo es significativo con un nivel de significación del 10%, del 5% y del 1%. Por último, en cuanto al signo de los estimadores de los parámetros de las variables que son significativos llegamos a las mismas conclusiones que las observadas en el modelo probit univariante.

4.2.2) Seguro de Automóvil

Respecto al seguro de automóvil, podemos observar en la [Tabla 8] como cuenta con once parámetros, de los cuales dos no son significativos en ninguno de los tres niveles de significación recogidos en la [Tabla 7] para ninguno de los dos modelos.

Si observamos el modelo probit univariante, el parámetro asociado a la variable tipo de automóvil sólo es significativo al 10%, mientras que los asociados a las variables sexo y número de asientos del automóvil son significativos al 10% y al 5%. Los parámetros asociados a las variables edad, edad del vehículo, mediador del seguro de automóvil, bonus/malus, antigüedad del seguro de automóvil y el término constante son significativos en los tres niveles de significación contrastados.

En cuanto al signo de los parámetros estimados asociados a las variables que son significativos, podemos ver como en cuatro de las variables cuantitativas éstos son significativos en algún nivel de significación de los tres antes comentados, habiendo relación directa. Así pues, a mayor edad del tomador, a mayor número de asientos, a mayor penalización (malus) y a mayor antigüedad de la póliza de automóvil, mayor será la probabilidad de renovación. En contraposición, cuando mayor sea la edad del vehículo la probabilidad de renovación disminuye.

Además, si tenemos en cuenta los parámetros asociados a las variables cualitativas que son significativos se observa que cuando el mediador del seguro de automóvil es un agente, las probabilidades de renovación aumentan, mientras que cuando el sexo es varón

y el tipo de auto es familiar las probabilidades de renovación disminuyen respecto al sexo mujer y al resto de usos no familiares.

Por otra parte, si observamos el modelo logit univariante, llegamos a las mismas conclusiones que en el modelo probit univariante en lo que a significación de los parámetros asociados a las variables se refiere, con la única diferencia que en el caso que nos concierne el parámetro de la variable capital contenido es significativo con un nivel de significación del 10% y del 5%. Por último, en cuanto al signo de los estimadores de los parámetros vinculados a las variables que son significativos llegamos a las mismas conclusiones que las observadas en el modelo probit univariante.

4.3) Modelización bivariante

Una vez hemos modelizado la cartera con los dos modelos univariantes, y habiendo dependencia entre las dos variables explicadas tal y como hemos visto con la tabla de contingencia, es el momento de obtener y analizar los resultados que nos puede proporcionar un modelo bivariante como es el probit bivariante.

Variables Explicativas / Modelos		**Hogar** **Probit Bivariante**	**Automóvil** **Probit Bivariante**
Comunes	**Intercept**	1,005251 ***	1,067231 ***
	Sexo	-0,013218	-0,09269 ***
	Edad	0,002666 *	0,00429 ***
Hogar	**Tipo Hogar**	-0,078523 *	
	Capital Continente	-0,000177	
	Capital Contenido	0,001393 *	
	Mediador Hogar	0,131854 ***	
	Forma de Pago Hogar	0,252578 ***	
	Antigüedad Hogar	0,009254 **	
Auto	**Tipo Auto**		-0,078187 *
	Potencia Auto		0,053162
	Edad Auto		-0,032012 ***
	Asiento Auto		0,045207 *
	Mediador Auto		0,225946 ***
	Bonus/Malus Auto		0,00056 ***
	Forma de Pago Auto		0,030314
	Antigüedad Auto		0,013826 ***
Término de correlación "Rho"		0,550876 ***	

[Tabla 9] Estimadores y contrastes de significación de los modelos probit bivariantes. Fuente: elaboración propia.

El modelo probit estimado en este apartado, tal y como se ha comentado con anterioridad, es un modelo bivariante que tiene en cuenta que el tomador se puede ver afectado por el hecho de haber renovado un seguro al renovar el otro.

En la [Tabla 9] se recoge, de manera análoga que con los modelos univariantes, los estimadores y los contrastes de significatividad para los niveles de significación de la [Tabla 7], que son el 10%, el 5% y el 1%, respectivamente.

En dicha [Tabla 9] podemos observar como la significación de los parámetros, tanto para el seguro de hogar como para el de automóvil, es muy parecida a la que veíamos en los modelos univariantes, aunque con pequeños cambios. Es por eso que para analizar la significación de los parámetros y el signo de los estimadores se compara el presente modelo con el modelo probit univariante, ya que el modelo probit bivariante se puede entender como dos modelos probit univariantes dependientes.

En cuanto a la significación de los parámetros del seguro de hogar, podemos observar como los únicos cambios respecto el modelo probit univariante se dan con las variables edad y tipo de hogar, en que en este caso los parámetros estimados pasan de no ser significativos a ser significativos sólo con un nivel de significación del 10%. Además, el parámetro asociado a la variable antigüedad de la póliza de hogar pasa a ser significativo sólo con niveles de significación del 10% y el 5%, siendo antes significativo en los tres niveles.

Si analizamos los signos de los estimadores, podemos ver como la interpretación de los estimadores significativos en el modelo probit univariante es la misma. Además, contamos con dos parámetros más que también son significativos. La variable edad es una variable cuantitativa que tiene una relación directa, lo que nos indica que a mayor edad mayor probabilidad de renovación. Por el contrario, la variable tipo de hogar es una variable cualitativa, la cual tiene una relación inversa, lo que nos indica que cuando el tipo de vivienda es un piso las probabilidades de retención disminuyen en comparación con las viviendas unifamiliares.

Por otro lado, si analizamos la significación de los parámetros del seguro de automóvil, podemos observar que se mantienen los mismos resultados, aunque con pequeños cambios en el nivel de significación. El parámetro asociado a la variable sexo pasa a ser significativo en los tres niveles de significación y, además el parámetro asociado a la variable número de asientos del vehículo pasa a ser significativo sólo con un nivel de significación del 10%.

En referencia a los signos de los estimadores, podemos observar como al haber los mismos parámetros significativos que en el caso univariante, en este caso también se interpretan los estimadores de los parámetros de manera idéntica a como se ha hecho en el modelo probit univariante.

4.4) Comparativa de los modelos

Considerando que la base de datos ha sido modelizada con modelos univariantes y bivariantes y que ya se han analizado los signos de los estimadores de los parámetros y su significación, es el momento de proceder a la comparación entre los tres modelos.

Primeramente se contrastan los modelos univariantes entre sí, para luego proceder a su comparación con el modelo probit bivariante y, así, ver si realmente este último es el que mejor se ajusta a los datos de nuestra cartera.

Modelos / Cálculos	-2logL Hogar	-2logL Automóvil	k	AIC
LOGIT	4856,0480	6539,5080	20	11435,556
PROBIT	4857,0430	6538,1800	20	11435,223

[Tabla 10] Cálculo del criterio de información de Akaike para los modelos univariantes. Fuente: elaboración propia.

En la [Tabla 10] se muestran los valores de los diferentes elementos que forman parte de la formulación utilizada para el cálculo del criterio de información de Akaike para los modelos univariantes. En el caso que tuviéramos que elegir un modelo univariante entre los dos usados, seleccionaríamos el modelo probit univariante, ya que es el que cuenta con un AIC levemente inferior, y por tanto, el que mejor se ajusta de los dos.

Modelo / Cálculo	AIC
PROBIT BIV.	10972

[Tabla 11] Criterio de información de Akaike para el modelo bivariante. Fuente: elaboración propia.

Si comparamos ahora los datos obtenidos en la [Tabla 10] con lo dispuesto en la [Tabla 11], se evidencia que el modelo probit bivariante se ajusta mejor a los datos reales y, por tanto, hace mejores predicciones que los modelos univariantes, puesto que su AIC es considerablemente menor que los de estos últimos.

Así, en el caso de que una empresa estimara estos modelos para conseguir la probabilidad de renovación estimada de los tomadores de su cartera de seguros de hogar y automóviles, debería tener en consideración los datos proporcionados por el modelo probit bivariante, ya que éste le ayudaría a tomar las medidas más adecuadas en cuanto a la retención y conservación de su clientela.

4.5) La tasa de retención y la estimación de la cartera

La tasa de retención o la probabilidad de renovación es el porcentaje que nos indica qué cantidad de tomadores de nuestra cartera confía un año más en nuestra entidad aseguradora y sus servicios sobre el total.

Para poder determinar dicha tasa con nuestros modelos univariantes, debemos prestar atención a la tabla de clasificación, la cual nos indica en un cierto nivel de probabilidad cómo será la estimación del modelo.

Seguro	Modelo Univariante	Nivel de Probabilidad	Tabla de Clasificación								
			Correcto		Incorrecto		Porcentajes				
			Evento	No Evento	Evento	No Evento	Correcto	Sensibilidad	Especificidad	Falso Positivo	Falso Negativo
Seguro de Hogar	Logit	**0.930**	6103	303	364	3271	63,8	65,1	45,4	5,6	91,5
		0,935	4794	391	276	4580	51,6	**51,1**	**58,6**	5,4	92,1
		0,940	3515	467	200	5859	39,7	37,5	70	5,4	92,6
	Probit	**0,930**	6129	312	355	3245	64,1	65,4	46,8	5,5	91,2
		0,935	4813	396	271	4561	51,9	**51,3**	**59,4**	5,3	92
		0,940	3492	478	189	5882	39,5	37,3	71,7	5,1	92,5
Seguro de Automóvil	Logit	**0,895**	5442	552	490	3557	59,7	60,5	53	8,3	86,6
		0,900	4971	621	421	4028	55,7	**55,2**	**59,6**	7,8	86,6
		0,905	4441	689	353	4558	51,1	49,3	66,1	7,4	86,9
	Probit	**0,890**	5798	513	529	3201	62,9	64,4	49,2	8,4	86,2
		0,895	5354	564	478	3645	58,9	**59,5**	**54,1**	8,2	86,6
		0,900	4901	640	402	4098	55,2	54,5	61,4	7,6	86,5

[Tabla 12] Tabla de Clasificación para los modelos univariantes. Fuente: elaboración propia.

Si observamos la [Tabla 12], en la primera columna tenemos el nivel de probabilidad, el cual a medida que se avanza para abajo (más nivel de probabilidad) va clasificando mejor los no eventos pero peor los eventos.

En la siguiente columna encontramos los eventos y no eventos correctos (evento quiere decir que el tomador renueva y no evento que no renueva), mientras que en la siguiente vemos los eventos y no eventos incorrectos. Es decir, los eventos y no eventos de la columna "Correcto" nos indica que para ese nivel de probabilidad hay un cierto número de renovaciones y anulaciones correctas, mientras que para la columna "Incorrecto" nos indica el número de renovaciones y anulaciones que no son tales, o lo que es lo mismo, que son anulaciones y renovaciones respectivamente.

La columna "Porcentaje Correcto" nos indica el porcentaje de aciertos, es decir, que si observamos la [Tabla 12] vemos que para un modelo logit univariante aplicado a la subcartera de hogar, con un nivel de probabilidad del 93%, seis mil ciento tres eventos más trescientos tres no eventos de un total de diez mil cuarenta y uno suponen un acierto del 63.8%.

A continuación la columna “Sensibilidad” nos proporciona el porcentaje de eventos bien clasificados (correctos), mientras que la columna “Especificidad” nos da el porcentaje de no eventos bien clasificados (correctos).

Por último, las columnas “Falso Positivo” y “Falso Negativo” nos proporcionan el porcentaje de eventos que realmente no lo son y el porcentaje de no eventos que realmente sí que lo son respectivamente, es decir, el porcentaje de veces que se ha estimado que un tomador renovaba y no ha sido así, y a la inversa.

Para determinar la tasa de retención de los seguros de hogar y de automóviles de nuestra cartera a partir de modelos univariantes nos interesan especialmente las columnas “Sensibilidad” y “Especificidad”. El nivel de probabilidad para el que dichas columnas estén más balanceadas al 50% es el nivel de probabilidad que nos proporciona un porcentaje de eventos (renovaciones) y no eventos (anulaciones) bien clasificados lo más equilibrado posible y, por tanto, es nuestra tasa de retención.

Si observamos la [Tabla 12], concretamente las columnas “Sensibilidad” y “Especificidad”, podemos ver marcados en negrita los porcentajes que son más equilibrados para cada modelo univariante y tipo de seguro.

Así pues, de acuerdo con la [Tabla 13], para el seguro de hogar tanto el modelo logit univariante como el probit univariante coinciden en que la tasa de retención para la subcartera de dichos seguros es del 93,5%. Por otro lado, para el seguro de automóvil el modelo logit univariante nos indica que la tasa de retención es del 90%, mientras que para el modelo probit univariante dicha tasa es del 89,5%. Como el equilibrio entre “Sensibilidad” y “Especificidad” para el nivel de probabilidad del 90% del modelo probit univariante también es muy correcto, determinamos pues que la tasa de retención de la subcartera de los seguros de automóvil es del 90%.

Tasa / Seguro	**Hogar**	**Automóvil**
Retención	93,5%	90%

[Tabla 13] Tasa de Retención para cada tipo de seguro. Fuente: elaboración propia.

No es necesario llevar a cabo el cálculo de las dos tasas de retención con el modelo bivariante, ya que, en ese caso, deberían ser calculadas a partir de las probabilidades marginales estimadas en el modelo bivariante, es decir, con las probabilidades univariantes para cada tipo de seguro y, por tanto, el resultado obtenido sería aproximadamente el mismo.

Si queremos realizar unas tablas de clasificación donde comparar los datos estimados con los modelos univariantes y bivariante con los datos reales, dependiendo del punto de corte que se determine habrá un mayor o menor porcentaje de especificidad y sensibilidad.

Como tanto con los modelos univariantes como con el bivariante hemos estimado las probabilidades de renovación individuales para cada tipo de seguro, dichas tasas de retención pueden ser útiles como puntos de corte, si se quiere crear unas tablas de clasificación donde comparar las estimaciones de las renovaciones y anulaciones de la cartera con los datos reales con una sensibilidad y especificidad equilibrada.

El funcionamiento es sencillo; si la probabilidad de renovación del individuo es superior o igual al punto de corte, que en nuestro caso es la tasa de retención, predeciremos que dicho individuo renovará la póliza, mientras que si es inferior, la predicción sería que anulará la misma.

A continuación se analizan y comparan las sensibilidades y especificidades de cada estimación.

4.5.1) Estimación univariante

La estimación univariante no tiene en cuenta la decisión de renovación o anulación de cada tomador en el otro seguro, ya que estos modelos solo recogen los efectos ejercidos por las variables explicativas y, por tanto, los resultados mostrados a continuación no son tan realistas como los del modelo bivariante.

En lo que a seguro de hogar respecta, a continuación podemos ver dos tablas que comparan los datos reales de la cartera con la estimación realizada en cada caso.

Probit Univariante Seguro de Hogar		Estimación		
		No Renueva	**Renueva**	**Total**
Datos Reales	**No Renueva**	399	268	667
		3,97	2,67	6,64
		59,82	40,18	
		8,06	5,27	
	Renueva	4554	4820	9374
		45,35	48	93,36
		48,58	51,42	
		91,94	94,73	
	Total	4953	5088	10041
		49,33	50,67	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 14] Estimación de la cartera de hogar con el modelo probit univariante. Fuente: elaboración propia.

Logit Univariante Seguro de Hogar		Estimación		
		No Renueva	Renueva	Total
Datos Reales	No Renueva	400	267	667
		3,98	2,66	6,64
		59,97	40,03	
		8,05	5,27	
	Renueva	4570	4804	9374
		45,51	47,84	93,36
		48,75	51,25	
		91,95	94,73	
	Total	4970	5071	10041
		49,5	50,5	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 15] Estimación de la cartera de hogar con el modelo logit univariante. Fuente: elaboración propia.

En la [Tabla 14] podemos ver como para el modelo probit univariante la sensibilidad de la predicción es del 51,42%, mientras que la especificidad es del 59,82%.

Que la sensibilidad sea del 51,42% quiere decir que nuestro estimador detecta que renovarán el 51,42% de los clientes que realmente renovarán, mientras que la especificidad sea del 59,82% quiere decir que nuestro modelo estima que anularán el 59,82% de los clientes que realmente anularán la póliza del seguro de hogar.

En la [Tabla 15] se recoge para el modelo logit univariante una sensibilidad del 51,25%, mientras que la especificidad es del 59,97%. Por tanto, podemos decir que para un mismo punto de corte y para el seguro de hogar, el modelo probit univariante estima un mayor porcentaje de individuos que realmente renovarán, mientras que el modelo logit univariante especifica un mayor porcentaje de individuos que realmente anularán la póliza.

Por otro lado, los datos obtenidos con los modelos univariantes para los seguros de automóvil son los siguientes.

Probit Univariante Seguro de Automóvil		Estimación		
		No Renueva	Renueva	Total
Datos Reales	No Renueva	645	397	1042
		6,42	3,95	10,38
		61,9	38,1	
		13,62	7,48	
	Renueva	4090	4909	8999
		40,73	48,89	89,62
		45,45	54,55	
		86,38	92,52	
	Total	4735	5306	10041
		47,16	52,84	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 16] Estimación de la cartera de automóviles con el modelo probit univariante. Fuente: elaboración propia.

Logit Univariante Seguro de Automóvil		Estimación		
		No Renueva	**Renueva**	**Total**
Datos Reales	**No Renueva**	634	408	1042
		6,31	4,06	10,38
		60,84	39,16	
		13,62	7,57	
	Renueva	4020	4979	8999
		40,04	49,59	89,62
		44,67	55,33	
		86,38	92,43	
	Total	4654	5387	10041
		46,35	53,65	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 17] Estimación de la cartera de automóviles con el modelo logit univariante. Fuente: elaboración propia.

En la [Tabla 16] podemos apreciar como para el modelo probit univariante la sensibilidad es del 54,55%, representando un total de cuatro mil novecientos nueve casos estimados como renovación que realmente son renovaciones de la póliza de un total de ocho mil novecientas noventa y nueve renovaciones. Por otro lado, el porcentaje de especificidad es del 61,9%, significando que se han estimado seis cientos cuarenta y cinco anulaciones correctamente de un total de mil cuarenta y dos casos.

A su vez, en la [Tabla 17] vemos como para el modelo logit univariante la sensibilidad es del 55,33%, es decir, superior que en el modelo probit univariante, mientras que la especificidad es del 60,84%, siendo en este caso inferior que en el modelo probit univariante.

4.5.2) Estimación bivariante

Después de comparar las estimaciones de los modelos univariantes con los datos reales, es el momento de realizar la comparativa entre los datos reales y el modelo probit bivariante, el cual además de recoger los efectos ejercidos por las variables explicativas, también recoge las dependencias entre las variables explicadas. Al recoger dichos efectos, las sensibilidades y especificidades obtenidas para un mismo punto de corte que en los casos univariantes pueden variar.

A continuación, se muestran los datos estimados para el seguro de hogar.

Probit Bivariante Seguro de Hogar		Estimación		
		No Renueva	Renueva	Total
Datos Reales	No Renueva	387	280	667
		3,85	2,79	6,64
		58,02	41,98	
		8,05	5,35	
	Renueva	4422	4952	9374
		44,04	49,32	93,36
		47,17	52,83	
		91,95	94,65	
	Total	4809	5232	10041
		47,89	52,11	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 18] Estimación de la cartera de hogar con el modelo probit bivariante. Fuente: elaboración propia.

En la [Tabla 18] observamos cómo del modelo probit bivariante se obtiene una sensibilidad del 52,83%, la cual es más alta que la observada en los modelos univariantes. Esto quiere decir que este modelo estima un mayor número de renovaciones bien identificadas.

En contraposición, la especificidad recogida en dicha tabla es del 58,02%, la cual es menor que la estimada en los modelos univariantes. Esto quiere decir que este modelo estima un menor número de anulaciones correctas para el mismo punto de corte que en los modelos univariantes.

A continuación, se muestran los datos estimados para el seguro de automóvil.

Probit Bivariante Seguro de Automóvil		Estimación		
		No Renueva	Renueva	Total
Datos Reales	No Renueva	637	405	1042
		6,34	4,03	10,38
		61,13	38,87	
		13,41	7,65	
	Renueva	4112	4887	8999
		40,95	48,67	89,62
		45,69	54,31	
		86,59	92,35	
	Total	4749	5292	10041
		47,3	52,7	100

Frecuencia
Porcentaje
% Fila
% Columna

[Tabla 19] Estimación de la cartera de automóviles con el modelo probit bivariante. Fuente: elaboración propia.

En la [Tabla 19] vemos que la sensibilidad observada para el mismo punto de corte que en el modelo univariante es del 54,32%, mientras que la especificidad es del 61,13%.

En este caso, la sensibilidad es menor que en los dos casos univariantes mientras que la especificidad es más alta que en el caso logit univariante y menor que en el caso probit univariante.

En las tablas anteriores se han mostrado los casos en los que la sensibilidad y la especificidad están equilibradas, es decir, cuando el punto de corte es la tasa de retención, pero si quisiésemos aumentar el nivel de sensibilidad o de especificidad deberíamos jugar con el punto de corte hasta encontrar la sensibilidad o especificidad deseada. En nuestro caso, si queremos obtener una estimación lo más correcta posible para las renovaciones deberíamos intentar aumentar el nivel de la sensibilidad de la estimación lo máximo posible, aceptando las posibles pérdidas implícitas de especificidad.

Elegir el modelo que mejor se ajusta a los datos reales, analizar los parámetros de las variables tanto por su signo como por su significación, estimar la tasa de retención y jugar con diferentes puntos de corte para ajustar la estimación del modelo elegido a nuestras necesidades son los pasos esenciales para conseguir entender y predecir el comportamiento de nuestros clientes.

Es por eso que, si se parte del trabajo realizado en el presente texto, se podrían plantear diferentes escenarios para ver cómo se comportarían los clientes y cómo variaría la tasa de retención y, a partir de aquí formalizar políticas y acciones que potencien la satisfacción del cliente en la compañía y, por tanto, su retención.

5) Conclusiones

Retener un cliente supone un coste mucho menor para una entidad aseguradora que convencer uno nuevo, por lo que las entidades han tenido que desarrollar diferentes herramientas que les permitan entender qué busca el cliente, cómo hacer que éste esté satisfecho con el servicio prestado y, además, entender qué motivos decantan su decisión de renovar o no renovar un año más sus seguros de hogar o de automóvil.

Aquí es donde actúan los modelos de retención de clientes. Estos modelos nos proporcionan la información necesaria para saber cómo afectan las diferentes variables de cada tomador en su elección de renovar o anular la póliza a fin de año.

En este texto se han empleado y analizado dos tipos de modelos diferentes; los modelos univariantes y los bivariantes. Los primeros estiman las probabilidades de renovación de cada asegurado sólo teniendo en cuenta cómo afectan sus variables explicativas, mientras que los segundos, no solamente recogen los efectos de las variables explicativas, sino que además también recogen la dependencia entre las variables dependientes. O lo que es lo mismo, cómo afecta sobre la decisión de renovar la póliza de hogar que ese tomador renueve la póliza de automóvil y viceversa.

A priori, puede parecer lógico que exista relación entre las variables explicadas, ya que si un asegurado tiene más de una póliza en una misma compañía, puede ser que le cueste más cambiar de compañía que a una persona que sólo tiene una póliza. Para poder comprobar si esto es cierto, se ha hecho un contraste de existencia de dependencia entre las dos variables explicadas, rechazándose la hipótesis nula de independencia.

Una vez se ha comprobado que existe dependencia entre las variables dependientes, es el momento de elegir qué modelo es el que mejor nos explica el comportamiento de los asegurados. Para comparar los diferentes modelos, se ha procedido a calcular el criterio de información de Akaike. Dicho criterio ha concluido que el modelo que mejor se ajusta a los datos modelizados es el probit bivariante.

Estos modelos nos permiten interpretar la relación existente entre las variables explicativas y las explicadas mediante el signo de sus parámetros estimados, además de contrastar si dichos parámetros son significativos a diferentes niveles de significación.

A modo de ejemplo, si tomamos como referencia el modelo probit bivariante, podemos sacar conclusiones válidas para ambos tipos de seguros, como que a mayor antigüedad de la póliza mayor probabilidad de renovar, cómo que si el mediador de seguros es un agente de la compañía, mayor será la probabilidad de renovación del cliente, o bien que los

clientes que pagan su seguro anualmente tienen más probabilidad de renovar con la compañía. También podemos identificar relaciones particulares de cada tipo de seguro, como que para un nivel de significación del 10%, si el tipo de vivienda es un piso, la probabilidad de renovación de la póliza de hogar disminuye, o bien que si el tipo de automóvil es familiar, la probabilidad de renovación también disminuye.

Por otro lado, también podemos comprobar que hay parámetros que son significativos para un tipo de seguro y para otro no; en los seguros de hogar el parámetro asociado a la variable sexo no es significativo mientras que sí que lo es en el seguro de automóvil.

Además, si se elabora una tabla de clasificación para el modelo utilizado se puede analizar para puntos de corte diferentes cómo será la sensibilidad y la especificidad de la estimación y, por tanto, calibrar dicha metodología acorde con nuestros objetivos.

Es por todo esto que, después de haber realizado este trabajo, he entendido por qué los modelos de retención de clientes son imprescindibles para las entidades aseguradoras hoy en día. Los modelos de retención de clientes son una herramienta de gran valor para las aseguradoras, ya que, si éstas necesitan modelos de tarificación para cuantificar el riesgo y ponerles un precio lo más atractivo posible para captar al cliente, también necesitan los modelos de retención de clientes para detectar porqué sus asegurados anulan, elaborar planes de actuación para mejorar la retención en su cartera, calcular tasas de retención condicionadas y plantear diferentes escenarios futuros para ver cómo variaría su tasa de retención.

Además, después de utilizar modelos univariantes y bivariantes, tengo claro que si alguna vez necesito realizar un modelo de retención de clientes en mi compañía, usaría un modelo bivariante o multivariante (dependiendo del número de variables dependientes), ya que es importante captar los posibles efectos que pueda ejercer la decisión de renovar o anular una póliza de un tipo de seguro en las demás.

Bibliografía

Bolancé, C., Guillen, M., y Padilla-Barreto, A.E. (2016a). "Predicting Probability of Customer Churn in Insurance". *Modeling and Simulation in Engineering, Economics and Management*, 82-91. Springer International Publishing.

Bolancé, C., Guillen, M., y Padilla-Barreto, A.E. (2016b). "Predicting Defection in Non-life Motor and Home Insurance". *Department of Econometrics, Riskcenter-IREA, University of Barcelona* (España).

Cameron, A. y Trivedi, P. (2005). *"Microeconometrics; Methods and Applications".* Cambridge University Press. New York (Estados Unidos).

Frees, E.W., Derrig, R.A., & Meyers, G. (2014). *"Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques"*.

Frees, E. W. (2009). *"Regression modeling with actuarial and financial applications"*. Cambridge University Press. (Reino Unido).

Guillen, M., Nielsen, J.P., Scheike, T.H., Pérez-Marín, A.M. (2012). "Time-varying effects in the analysis of customer loyalty: a case study in insurance". *Expert Syst. Appl. 39(3), 3551-3558.*

Greene, W. (1999). *"Análisis Econométrico".* Prentice Hall. Madrid (España).

Wilson, J. (2009). "An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression". *Journal of Finance and Accountancy*.

Takeshi, A. (1985). *"Advanced Econometrics".* Basil Blackwell. Oxford (Reino Unido).

Directiva Oficial de la Unión Europea (2004). "*DIRECTIVA DEL CONSEJO 2004/113/CE de 13 de diciembre de 2004 por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro".* Diario Oficial de la Unión Europea L 373/37 del 21 de Diciembre de 2004 (España).

Anexo: Código en SAS

A continuación se muestra el código utilizado para realizar la parte práctica del trabajo.

Los programas utilizados para llevar a cabo estas modelizaciones, así como para elaborar los gráficos y tablas ilustrados en el presente trabajo son el Microsoft Excel y el SAS - tanto el SAS Enterprise Guide como el SAS University Edition -.

/******** 4) APLICACIÓN PRÁCTICA: MODELIZACIÓN Y ANÁLISIS ********/

/* Indicamos la librería que vamos a usar */

```
libname k "F:\TFM\SAS\Tablas SAS";
```

/* Importamos los datos de la cartera */

```
data k.datos;
infile "F:\TFM\SAS\datos_sas.csv" firstobs=2 dlm=";" ;
length
sexo $5.
edad 3.
estado_hogar $1.
tipo_hogar $2.
cap_continente_1 8.
cap_contenido_1 8.
mediador_hogar $8.
forma_pago_hogar $1.
antig_hogar 3.
estado_auto $1.
tipo_auto $20.
potencia_auto 3.
edad_auto 3.
asientos_auto 3.
mediador_auto $8.
bonus_malus_auto 8.
forma_pago_auto $1.
```

```
antig_auto 3.
;
input
sexo
edad
estado_hogar
tipo_hogar
cap_continente_1
cap_contenido_1
mediador_hogar
forma_pago_hogar
antig_hogar
estado_auto
tipo_auto
potencia_auto
edad_auto
asientos_auto
mediador_auto
bonus_malus_auto
forma_pago_auto
antig_auto
;
informat
cap_continente_1 commax20.2
cap_contenido_1 commax20.2
potencia_auto commax20.2
bonus_malus_auto commax20.2
;
format
cap_continente_1 commax20.2
cap_contenido_1 commax20.2
potencia_auto commax20.2
bonus_malus_auto commax20.2
;
run;


/* Dicotomizamos las variables carácter */

data k.trat_datos_1;
set k.datos;
if sexo="Mujer" then sexo=0; /* 0 quiere decir Mujer */
else if sexo="Varon" then sexo=1; /* 1 quiere decir Hombre */
if estado_hogar="V" then estado_hogar=1; /* 1 quiere decir que renueva */
```

```
else if estado_hogar="A" then estado_hogar=0; /* 0 quiere decir que anula y por tanto no renueva */
if tipo_hogar in ("UA" "UF" "RU") then tipo_hogar=0; /* 0 quiere decir Finca */
else if tipo_hogar in ("AT" "PB" "PI") then tipo_hogar=1; /* 1 quiere decir Piso */
if forma_pago_hogar="A" then forma_pago_hogar=1; /* 1 quiere decir que el pago es anual */
else if forma_pago_hogar in ("T" "S") then forma_pago_hogar=0; /* 0 quiere decir que el pago es fraccionado */
if mediador_hogar="Corredor" then mediador_hogar=0; /* 0 quiere decir que el mediadior es un corredor */
else if mediador_hogar="Agente" then mediador_hogar=1; /* 1 quiere decir que el mediadior es un agente */
if estado_auto="V" then estado_auto=1; /* 1 quiere decir que renueva */
else if estado_auto="A" then estado_auto=0; /* 0 quiere decir que anula y por tanto no renueva */
if tipo_auto in ("Berlina 2 volumenes" "Berlina 3 Volumenes" "Familiar") then tipo_auto=1; /* 1 quiere decir que tiene un uso familiar */
else if tipo_auto in ("Cabrio" "Coupe" "Turismo Pick Up" "NA") then tipo_auto=0; /* 0 quiere decir que tiene un uso diferente del familiar */
if forma_pago_auto="A" then forma_pago_auto=1; /* 1 quiere decir que el pago es anual */
else if forma_pago_auto in ("T" "S") then forma_pago_auto=0; /* 0 quiere decir que el pago es fraccionado */
if mediador_auto="Corredor" then mediador_auto=0; /* 0 quiere decir que el mediadior es un corredor */
else if mediador_auto="Agente" then mediador_auto=1; /* 1 quiere decir que el mediadior es un agente */
run;
```

/* Convertimos en variables numéricas las variables carácter ya dicotomizadas */

```
data k.trat_datos_2;
set k.trat_datos_1;
sexo2=input(sexo,best20.);
estado_hogar2=input(estado_hogar,best20.);
tipo_hogar2=input(tipo_hogar,best20.);
forma_pago_hogar2=input(forma_pago_hogar,best20.);
mediador_hogar2=input(mediador_hogar,best20.);
estado_auto2=input(estado_auto,best20.);
tipo_auto2=input(tipo_auto,best20.);
forma_pago_auto2=input(forma_pago_auto,best20.);
mediador_auto2=input(mediador_auto,best20.);
run;
```

```
data k.trat_datos_3;
set k.trat_datos_2;
drop sexo estado_hogar tipo_hogar forma_pago_hogar mediador_hogar estado_auto tipo_auto forma_pago_auto mediador_auto;
run;

data k.trat_datos_4;
set k.trat_datos_3;
rename sexo2=sexo estado_hogar2=estado_hogar tipo_hogar2=tipo_hogar forma_pago_hogar2=forma_pago_hogar mediador_hogar2=mediador_hogar
estado_auto2=estado_auto tipo_auto2=tipo_auto forma_pago_auto2=forma_pago_auto mediador_auto2=mediador_auto;
run;


/* Dividimos entre 1.000 las variables capital "cap_continente_1" y "cap_contenido_1"*/

data k.trat_datos_5;
set k.trat_datos_4;
cap_continente=cap_continente_1/1000;
cap_contenido=cap_contenido_1/1000;
drop cap_continente_1 cap_contenido_1;
run;


/* Ordenamos las variables */

proc sql;
create table k.trat_datos as
select sexo, edad, estado_hogar, tipo_hogar, cap_continente, cap_contenido, mediador_hogar, forma_pago_hogar, antig_hogar, estado_auto,
tipo_auto, potencia_auto, edad_auto, asientos_auto, mediador_auto, bonus_malus_auto, forma_pago_auto, antig_auto
from k.trat_datos_5;
run;


/* ESTADÍSTICOS DESCRIPTIVOS */


/* Variables CUALITATIVAS */
```

```
proc sql;
create view ordenar as
select t.sexo, t.estado_hogar, t.tipo_hogar, t.mediador_hogar, t.forma_pago_hogar,
t.estado_auto, t.tipo_auto, t.mediador_auto, t.forma_pago_auto
from k.trat_datos as t;
quit;

proc freq data=ordenar order=internal;
tables sexo / scores=table;
tables estado_hogar / scores=table;
tables tipo_hogar / scores=table;
tables mediador_hogar / scores=table;
tables forma_pago_hogar / scores=table;
tables estado_auto / scores=table;
tables tipo_auto / scores=table;
tables mediador_auto / scores=table;
tables forma_pago_auto / scores=table;
run;


/* Variables CUANTITATIVAS */

proc sql;
create view ordenar_2 as
select t.edad, t.cap_continente, t.cap_contenido, t.antig_hogar, t.potencia_auto,
t.edad_auto, t.asientos_auto, t.bonus_malus_auto, t.antig_auto
from k.trat_datos as t;
quit;

proc means data= ordenar_2 fw=12
printalltypes chartype vardef=df
mean std stderr var min max mode range sum n nmiss;
var edad cap_continente cap_contenido antig_hogar potencia_auto edad_auto
asientos_auto bonus_malus_auto antig_auto;
run;


/* 4.1) TABLA DE CONTINGENCIAS DE LAS VARIABLES DEPENDIENTES */

proc freq data=k.trat_datos;
tables estado_hogar*estado_auto / chisq;
run;
```

/* 4.2) MODELIZACIÓN UNIVARIANTE */

/* Usamos descending para que modelize la prob de (Y=1) */

/* HOGAR */

/* Modelo PROBIT */

```
proc logistic data=k.trat_datos descending;
model estado_hogar = sexo edad tipo_hogar cap_continente cap_contenido mediador_hogar forma_pago_hogar antig_hogar /ctable pprob= (0.92 to 0.96 by 0.005) link=probit;
output out=k.trat_datos_predh1 pred=pred;
run;

data k.trat_datos_pred_h1;
 set k.trat_datos_predh1;
 Y_pred=1;
 if pred<0.935 then Y_pred=0;
 run;

proc freq data=k.trat_datos_pred_h1;
table estado_hogar*Y_pred;
run;
```

/* Modelo LOGIT */

```
proc logistic data=k.trat_datos descending;
model estado_hogar = sexo edad tipo_hogar cap_continente cap_contenido mediador_hogar forma_pago_hogar antig_hogar /ctable pprob= (0.92 to 0.94 by 0.005) link=logit;
output out=k.trat_datos_predh2 pred=pred;
run;

data k.trat_datos_pred_h2;
 set k.trat_datos_predh2;
 Y_pred=1;
 if pred<0.935 then Y_pred=0;
 run;

proc freq data=k.trat_datos_pred_h2;
table estado_hogar*Y_pred;
run;
```

/* AUTOMÓVILES */

/* Modelo PROBIT */

```
proc logistic data=k.trat_datos descending;
model estado_auto = sexo edad tipo_auto potencia_auto edad_auto asientos_auto mediador_auto bonus_malus_auto forma_pago_auto antig_auto /ctable pprob= (0.88 to 0.92 by 0.005) link=probit;
output out=k.trat_datos_preda3 pred=pred;
run;

data k.trat_datos_pred_a3;
 set k.trat_datos_preda3;
 Y_pred=1;
 if pred<0.9 then Y_pred=0;
 run;

proc freq data=k.trat_datos_pred_a3;
table estado_auto*Y_pred;
run;
```

/* Modelo LOGIT */

```
proc logistic data=k.trat_datos descending;
model estado_auto = sexo edad tipo_auto potencia_auto edad_auto asientos_auto mediador_auto bonus_malus_auto forma_pago_auto antig_auto /ctable pprob= (0.88 to 0.92 by 0.005) link=logit;
output out=k.trat_datos_preda4 pred=pred;
run;

data k.trat_datos_pred_a4;
 set k.trat_datos_preda4;
 Y_pred=1;
 if pred<0.9 then Y_pred=0;
 run;

proc freq data=k.trat_datos_pred_a4;
table estado_auto*Y_pred;
run;
```

/* 4.3) MODELIZACIÓN BIVARIANTE: Modelo Probit Bivariante*/

/* HOGAR Y AUTOMÓVILES */

```
proc qlim data=k.trat_datos method=qn;
  model estado_hogar = sexo edad tipo_hogar cap_continente cap_contenido
mediador_hogar forma_pago_hogar antig_hogar;
  model estado_auto = sexo edad tipo_auto potencia_auto edad_auto asientos_auto
mediador_auto bonus_malus_auto forma_pago_auto antig_auto;
  endogenous estado_hogar estado_auto ~ discrete;
  output out=k.trat_datos_pred2 XBETA;
run;

data k.trat_datos_pred3;
 set k.trat_datos_pred2;
  prob1=probnorm(Xbeta_estado_hogar); *Mirar el nombre de xbeta1 y xbeta2;
  prob2=probnorm(Xbeta_estado_auto);
  Y1_pre=1;
  if prob1<0.935 then Y1_pre=0;
  Y2_pre=1;
  if prob2<0.9 then Y2_pre=0;
run;

proc freq data=k.trat_datos_pred3;
table estado_hogar*Y1_pre;
run;

proc freq data=k.trat_datos_pred3;
table estado_auto*Y2_pre;
run;
```