uc3m | Universidad Carlos III de Madrid

Máster Universitario en Ciencias Actuariales y Financieras
2020-2021

Trabajo Fin de Máster

Optimización matemática a partir de algoritmos híbridos.
Una aplicación en la tarificación del seguro de automóviles

Juan Sánchez Campillo

Tutores

José Miguel Rodríguez-Pardo del Castillo
Jesús Ramón Simón del Potro

Madrid, 2021

Firmado:

DETECCIÓN DEL PLAGIO

La Universidad utiliza el programa **Turnitin Feedback Studio** *para* comparar la originalidad del trabajo entregado por cada estudiante con millones de recursos electrónicos y detecta aquellas partes del texto copiadas y pegadas. Copiar o plagiar en un TFM es considerado una **Falta Grave**, y puede conllevar la expulsión definitiva de la Universidad.

RESUMEN

En los últimos años ha surgido con fuerza el uso de modelos predictivos de Inteligencia Artificial en todos los sectores. El ámbito actuarial no es una excepción. En la tarificación de primas algunas entidades se están planteando sustituir los métodos tradicionales de tarificación por estos nuevos algoritmos que ofrecen, con frecuencia, mayor capacidad predictiva. En este trabajo se discuten y evalúan las ventajas de cada enfoque y se propone usar una combinación de ambas técnicas en lo que se define como modelos híbridos, que combinan los métodos clásicos (GLM) con modelos de Inteligencia Artificial, para mejorar las predicciones sin perder la capacidad explicativa necesaria en una actividad regulada como es la aseguradora. Asimismo, se ha desarrollado un algoritmo de optimización matemática que maximiza el margen esperado individual, de cada póliza, y que modela no sólo el coste del riesgo, sino también el comportamiento del tomador frente a la prima ofrecida. El trabajo se acompaña con una aplicación práctica de los modelos a una cartera de pólizas del seguro de automóviles, desarrollada íntegramente en la estructura de aplicaciones web R Shiny.

Palabras clave: Inteligencia Artificial, *Machine Learning*, optimización individual de primas, precios basados en el comportamiento, seguro de autos.

ABSTRACT

In recent years, the use of artificial intelligence predictive models has developed sharply in all sectors. The actuarial field is no exception for this. In pricing, there are suggestions that these new algorithms will replace traditional pricing methods, offering greater predictive capabilities. In this paper, the advantages of each approach are discussed and evaluated, a combination of both classical methods (GLM) with Artificial Intelligence models are combined into hybrid models. These new algorithms improve predictions without losing the necessary explanatory capacity required in a regulated activity such as insurance. Likewise, a complete mathematical optimization algorithm has been developed that maximizes the individual expected margin of each policy, and which models not only the cost of the risk, but also the behaviour of the policyholder in relation to the premium offered. The work is accompanied by a practical application of the models to a portfolio of motor insurance policies, developed entirely in the R Shiny web application framework.

Keywords: Artificial Intelligence, Machine Learning, individual pricing optimization, behavioral pricing, motor insurance.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. MOTIVACIÓN

La aparición del automóvil a finales del siglo XIX supuso una revolución en la movilidad y las comunicaciones de las personas. Además de las evidentes consecuencias positivas que tuvo su nacimiento, pronto surgió la necesidad del aseguramiento como consecuencia de los accidentes.

Tal y como se refiere en Fernández (2016), existen distintas hipótesis sobre los orígenes del primer seguro de automóviles si bien hay coincidencia en el marco temporal de su inicio en la última década del siglo XIX. Por un lado, el Reino Unido se atribuye la creación en 1895 de una póliza específica de autos. A su vez Estados Unidos refiere la primera póliza el 1 de febrero de 1898. La hipótesis más extendida es que la compañía norteamericana *Traveller Insurance* vendió la primera póliza en Nueva York con una suma asegurada de 5.000 dólares y una prima de 12,25 dólares. Esta primera póliza cubría la Responsabilidad Civil derivada de accidentes, no con otros vehículos de motor y sí contra coches de caballos o jinetes.

La universalización de coche como principal instrumento de la movilidad de las personas, con el consiguiente incremento de las vías y de los vehículos en circulación, trajo consigo la necesidad del aseguramiento y de convertir, al menos en parte, ese aseguramiento en obligatorio.

El seguro de automóviles comenzó a desarrollarse inicialmente con coberturas de responsabilidad civil a la que fueron añadiéndose nuevas garantías aseguradoras relacionadas con los daños propios de los vehículos y con coberturas de asistencia al vehículo y a los conductores y pasajeros.

El numeroso parque de vehículos, junto con la obligatoriedad en la suscripción obligatoria de la responsabilidad civil, constituye el principal motivo por el que el seguro de automóviles es hoy en día el ramo principal de seguros de no vida en todo el mundo.

En la actualidad esta modalidad aseguradora se encuentra inmersa en un proceso de evolución, que apunta a una profunda transformación como se recoge en ICEA

(2018). Algunos de los cambios disruptivos que afectarán a este tipo de seguros son:

- La aparición de las tecnologías de ayuda a la conducción[1] que están reduciendo la frecuencia siniestral. La incorporación masiva de estas ayudas permitirá una reducción del precio del seguro y un incremento de la importancia de las variables asociadas al vehículo frente a aquellas relacionadas con las características de los conductores. El empleo de una adecuada clasificación de vehículos se convierte, por tanto, en un elemento fundamental en una adecuada tarificación de este seguro.
- El cambio en el tipo de combustión de los vehículos, con una traslación de los motores de combustión fósil, a coches con motorizaciones más sostenibles, como la eléctrica, que pueden suponer una modificación en los patrones de siniestralidad.
- La instalación de dispositivos embarcados[2] en los vehículos que permiten medir más granularmente el riesgo a través de variables como el kilometraje recorrido, las vías por la que se circula, la velocidad media o las aceleraciones del vehículo. Estos dispositivos proporcionan elevados volúmenes de información que permiten una valoración más exacta del riesgo. En Johnston (2020) se señala como la tecnología telemática embarcada permite una valoración de los conductores mucho más precisa que la usada hasta ahora. Esto se debe a que realiza una medición individual que va más allá que la utilizada tradicionalmente como la edad, el número de años del carné de conducir, la zona geográfica o el historial de siniestralidad.
- Los cambios en la conducta de los clientes ante la compra del seguro como consecuencia de la incorporación de las nuevas cohortes de población con una mayor propensión a la compra por internet.

[1] Conocida con el acrónimo anglosajón ADAS ("*Advanced Driver Assistance Systems*").

[2] Conocidos en la jerga por su acrónimo anglosajón, dispositivos OBD ("*On Board Diagnostics*").

- La entrada creciente de nuevos actores en el aseguramiento de los vehículos como los fabricantes de coches, los agregadores de precio o los grandes gigantes de internet (Google, Apple, Facebook o Amazon).
- Los cambios en la percepción de las necesidades de las personas en relación a la movilidad que conducirá a la coexistencia de vehículos en propiedad, con la consideración del vehículo como un servicio (modelos de *car sharing*).

En este entorno de incertidumbre sobre la evolución del ramo de automóviles nos encontramos adicionalmente con tomadores cada vez más informados. Además, el elevado grado de competencia entre compañías aseguradoras y la *comoditización*[3] del seguro provoca que la prima sea un elemento determinante para el éxito o el fracaso en la gestión del ramo de automóviles. En Anraoui et al. (2009) se indica además como esta falta de diferenciación lleva aparejada una pérdida del valor añadido que tradicionalmente aportaba el canal agencial en el asesoramiento en la contratación de seguros.

Aquellas entidades que sean capaces de emplear técnicas más sofisticadas para determinar sus tarifas, a partir de la mejora de los modelos actuariales de riesgo para medir la siniestralidad esperada y del análisis del comportamiento del cliente frente a la prima ofrecida, serán las que consigan mejores resultados en un negocio cada vez más competitivo.

En este contexto de mercado se hace imprescindible la utilización de modelos estadísticos avanzados, a través de unos algoritmos de tarificación que pueden llegar a ser más predictivos[4], como los que proporcionan las técnicas de *Machine Learning*.

[3] Se utiliza este anglicismo para referirse al contexto de mercado en el que se considera que no hay diferencias en el producto ofrecido, por lo que las diferencias sólo vienen marcadas por el precio.

[4] La capacidad predictiva se mide tanto en precisión como en robustez de la tarificación. Precisión en el sentido de calcular una prima pura más ajustada para cada asegurado, con un intervalo de confianza y unas tasas de error más reducidas. Robustez, para que las predicciones sean lo más estables posibles ante valores atípicos (Guillén y Pesantez-Narvaez, 2018).

Como elemento negativo a considerar se tiene que, frente al mayor carácter predictivo de los algoritmos, estos modelos presentan el inconveniente de la falta de trasparencia como se señala en Rodríguez-Pardo (2017). Para prevenir un uso ilícito de estos métodos, el desarrollo adecuado de las técnicas de Inteligencia Artificial debe convertirse en un objetivo a cumplir por parte de las entidades aseguradoras y, para vigilarlo, ya existen iniciativas en el ámbito privado como la creación de comités de ética en el uso de estas técnicas[5].

Desde una perspectiva supervisora, la falta de verificabilidad puede convertirse también en un problema[6] en un mercado altamente regulado, donde la expresión matemática de fijación de la prima debe incorporarse en las bases técnicas del producto, a través de fórmulas que permitan verificar que se cumple con la legislación. Así el artículo 118 del ROSSEAR[7] establece la necesidad de que se explicite la equivalencia actuarial en la fijación de la prima, todo ello para preservar los principios que equidad, no discriminación y suficiencia que debe tener la tarifa. Los GLMs[8] proporcionan una interpretabilidad directa e intuitiva, como indican Guillén y Pesantez-Narvaez (2018), lo que facilita medir el impacto de cada factor de riesgo en el cálculo de la prima y son fácilmente trasladables a las notas técnicas de los productos.

Adicionalmente los nuevos algoritmos presentan una mayor complejidad en la incorporación de los cálculos en los sistemas operacionales de gestión de las entidades aseguradoras comparados con los GLMs. En este sentido, los modelos

[5] Como el desarrollado por la Mutualidad de la Abogacía para vigilar el uso de la Inteligencia Artificial dentro del cumplimiento de los principios y valores de esa Entidad (Cinco Días, 2019).

[6] Aunque existen nuevas metodologías que permiten obtener tarificadores también en los algoritmos de *Machine Learning* (Witten et al., 2016).

[7] Real Decreto 1060/2015, de 20 de noviembre, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras.

[8] *Generalized linear models* o GLM por su acrónimo en inglés. Modelos lineares generalizados en castellano.

predictivos clásicos son fácilmente programables en los sistemas de emisión de las compañías.

En la literatura actuarial reciente, es frecuente enfrentar las técnicas tradicionales de tarificación, como los modelos lineales generalizados, con estas técnicas más novedosas de algoritmos de Inteligencia Artificial, comparando la capacidad predictiva de los distintos métodos (Zhou y Deng, 2019). En muchos casos se opta por la utilización alternativa de una técnica u otra en función del criterio de éxito seleccionado. No obstante, algunos nuevos enfoques, como el que sugiere Panlilio et al. (2018), proponen combinar las técnicas aprovechando los mejores atributos de cada una de ellas para resolver la cuestión actuarial planteada. En ese sentido, la integración de las nuevas técnicas de Inteligencia Artificial, junto con los algoritmos de cálculo utilizados tradicionalmente en la tarificación del seguro del automóvil, permiten la obtención y trazabilidad de la prima a través de una expresión algebraica, resolviendo por tanto el inconveniente de la falta de trasparencia sin renunciar a la mejora en la precisión en el cálculo de la prima.

En este trabajo se propone la utilización de estos modelos estadísticos híbridos para la mejora de la capacidad predictiva de las tarifas que ofrecen los modelos del tipo GLM. El trabajo tiene un enfoque integral que comprende no sólo la estimación del coste del riesgo, sino también la medición del comportamiento del cliente frente a la prima propuesta en la renovación de su póliza.

2. OBJETIVOS

2.1 Objetivo principal

El objetivo principal de este trabajo es proponer una combinación de algoritmos de *Machine Learning* y de modelos lineales generalizados para mejorar la capacidad predictiva de los resultados de la optimización individual de primas en el seguro del automóvil. Este uso conjunto de algoritmos tiene la ventaja adicional de que mantiene las propiedades deseables de los modelos clásicos, en lo referido a la sencillez en la implementación operativa e interpretabilidad, sin renunciar a la mejora de las predicciones que proporcionan los algoritmos de *Machine Learning*. A esta combinación de modelos se la conoce como modelos híbridos.

La combinación de algoritmos híbridos se empleará en los modelos que intervienen en el cálculo óptimo de la prima comercial, desde la derivación del precio técnico, hasta la estimación del precio de los competidores en la determinación de las curvas de demanda.

2.2 Objetivos secundarios

Además del objetivo principal se abordarán otros objetivos secundarios:

1. Descripción metodológica entre los modelos estadísticos clásicos y nuevas técnicas de *Machine Learning* analizando las ventajas e inconvenientes de cada uno de los métodos.

2. Realización de un *scoring* de vehículo con técnicas GLM y de *Machine Learning* y el análisis de su capacidad predictiva.

3. Análisis de las variables específicas del vehículo que se incorporarán a los modelos GLM de la prima de riesgo. Se profundizará en el efecto que tiene el

tipo de combustión del vehículo (gasolina/diésel/eléctrico) y la incorporación de nuevos dispositivos ADAS en la siniestralidad.

4. Realización de un modelo híbrido de prima pura.

5. Realización de un programa para la obtención de los precios de los competidores en internet a través de programación con técnicas de *web-scraping*.

6. Predicción de las tarifas de la competencia mediante el uso de GBMs.

7. Realización de un modelo híbrido de retención.

8. Aplicación de técnicas de optimización individual en la renovación de una cartera de automóviles y su comparación con técnicas tradicionales de renovación de precios.

9. Derivación y cálculo de todos los algoritmos referidos anteriormente a través de la utilización del software libre R.

3. ALCANCE

El alcance del trabajo se orienta al análisis de técnicas híbridas de tarificación aplicadas al seguro de automóviles en España, en suscripción individual usando técnicas de optimización de precios aplicadas a la renovación anual de las pólizas.

4. EL SEGURO DE AUTOMÓVILES EN ESPAÑA

Como en los países de nuestro entorno, los seguros de autos en España son los de mayor peso dentro del negocio no vida. Este es un negocio consolidado con cifras que se han mantenido durante los últimos años. Como resulta obvio el peso del negocio está muy condicionado por el total de vehículos en circulación por la necesidad de que estos cuenten con un seguro de responsabilidad civil obligatorio. En ese sentido, en el año 2019 el parque automovilístico en España alcanzó la cifra de 34,4 millones de vehículos, lo que supone un 2,1% más que el año anterior[9] aunque esta cifra se ha visto reducida en el año 2020 como consecuencia de la crisis económica surgida a raíz del COVID-19.

De acuerdo con las cifras sectoriales reportadas por ICEA (2021) el ramo de automóviles mantuvo una cuota de mercado del 30% alcanzando un volumen de primas superior a 11.000 millones de euros a diciembre del año 2020 lo que representa una disminución del 2% cifra muy semejante a la reducción del parque de vehículos. En cuanto a la tendencia del peso de este negocio frente al total de no vida presenta un decrecimiento continuado que le ha hecho perder 5,8 puntos porcentuales en los últimos diez años (MAPFRE ECONOMICS, 2020).

En la figura[10] que aparece a continuación se presenta el peso de la distribución por tipo de negocio considerando las primas de los seguros no vida en España.

[9] Extraído de (DGT, 2021).

[10] *Ibid.*

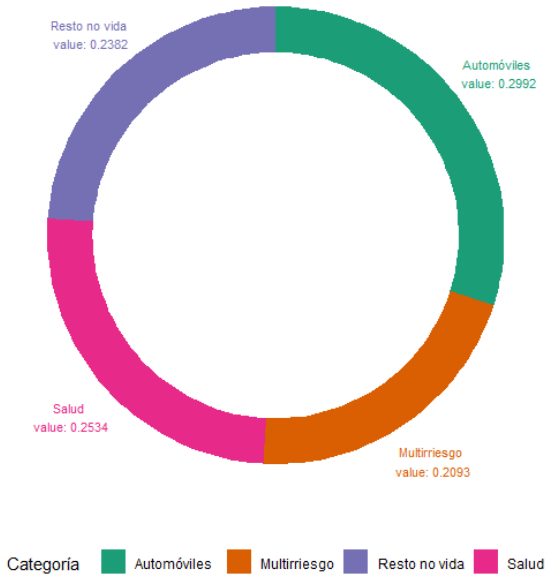


Figura 4.1: Distribución de primas de los seguros no vida

Del total de la distribución del negocio de automóviles, el mayor volumen se corresponde con la primera categoría[11] tal y como se refleja en el gráfico siguiente:

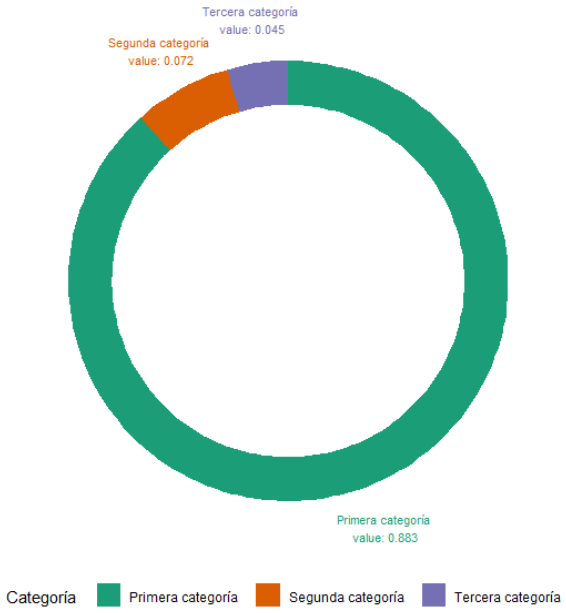


Figura 4.2: Distribución de primas de automóviles por categorías

[11] En España se clasifican los vehículos de motor en tres categorías. La primera categoría se refiere a automóviles, la segunda a camiones y autobuses y la tercera a motocicletas y ciclomotores.

La siniestralidad se ha visto reducida en el año 2020 como consecuencia del COVID-19, cuyas restricciones a la movilidad han provocado una disminución de los ratios técnicos. La cobertura de Retirada del Carnet, con una cifra de 57,7%, fue la que presentó una siniestralidad menor, mientras que la correspondiente a los autocares exhibió la mayor con un valor de un 68%. Adicionalmente, se produce una disminución tanto de la frecuencia como la severidad de los siniestros. Por coberturas, la frecuencia de daños propios es la que más se ha reducido con una variación del -13,9%. En lo relativo al coste medio las coberturas que más han bajado han sido Defensa Jurídica y Retirada del Carnet con una disminución del -28,2% y -26,7% respectivamente.

La modelización de prima que se propone en este estudio se corresponde, por tanto, con la modalidad aseguradora de no vida más relevante en la actualidad, y dentro de ella, con la primera categoría que es la que representa la mayoría del negocio. Las técnicas estadísticas a emplear que se proponen en este trabajo resultan especialmente adecuadas para la modelización de este negocio masa.

El comportamiento atípico de la siniestralidad en España durante el 2020 para todas las coberturas hace pertinente eliminar la información de este año para los métodos estadísticos de cálculo de prima que se proponen en este trabajo.

Además de los retos a los que se enfrenta el sector, enunciados en la introducción de este estudio, el negocio de automóviles se encuentra inmerso en una *guerra de precios* que se ha traducido en una disminución de la retención del negocio de 8,7 puntos durante los últimos 8 años (MAPFRE ECONOMICS, 2020). En este contexto de alta competencia, es imprescindible el empleo de técnicas avanzadas de modelización actuarial que traten de maximizar el resultado técnico y/o la retención del negocio suscrito.

5. TAXONOMÍA DE LA TARIFICACIÓN DEL SEGURO DE AUTOMÓVILES

El proceso de tarificación en los seguros constituye uno de los momentos clave del éxito o el fracaso de las entidades aseguradoras en su relación con los clientes.

Como ya se ha referido anteriormente, en relación al seguro de automóviles, nos encontramos con un mercado altamente competitivo, donde existe obligatoriedad en la suscripción de algunas coberturas y hay una *comoditización* del producto que lleva a unos altos niveles de competencia. Esto ha provocado que esta modalidad aseguradora se convierta, para dar respuesta a los requerimientos del mercado, en la más avanzada en términos de tarificación.

En los procesos de tarificación se pueden utilizar distintas aproximaciones al cálculo que dependen de distintas circunstancias que condicionan la metodología a elegir. Algunas de éstas, que pueden condicionar el modelo a elegir podrían ser el grado de competencia, el nivel de tecnificación del mercado o la legislación en materia aseguradora. En cualquier caso, tal y como se indica en Coskun (2016) o en Werner y Modlin (2016), se trata de resolver en todos los casos la siguiente ecuación:

$$Prima = Coste + Beneficio$$

Para la mayoría de productos no aseguradores el coste es conocido a priori. Sin embargo, en los seguros el coste no es conocido en el momento de la suscripción del producto. Este hecho dificulta y hace necesario la realización de modelos estadísticos para derivar el precio final.

Particularizando la fórmula anterior al mundo asegurador, se tendría la siguiente expresión:

$$Prima = Siniestralidad + Gastos + Beneficio$$

El proceso de tarificación consiste en derivar una prima en función de la siniestralidad, los gastos de adquisición y administración y el beneficio asignado. Este cálculo es prospectivo lo que significa que la tarifa pura es una estimación del valor futuro del coste que se deriva teniendo en cuenta la información histórica de los riesgos, así como otras variables que corrijan los datos históricos en función de los cambios futuros previstos, como por ejemplo la inflación, los cambios normativos o las variaciones en la composición de negocio (Werner y Modlin, 2016).

Un elemento clave en la cuantificación de la prima lo constituye la segmentación del riesgo según distintos niveles que pueden tener los factores utilizados. La adecuada segmentación de la prima permite seleccionar los mejores riesgos y descartar los peores discriminando los perfiles por precio. Por tanto, promueve la selección positiva y previene la antiselección.

En cuanto a la tipología de factores, en el seguro de automóviles se pueden encontrar tres grandes grupos de variables que permiten la discriminación de los riesgos:

1. Factores relativos al tomador, asegurado o a los conductores de la póliza. Edad, antigüedad del carnet, *credit scoring* o el nivel de *bonus-malus* son algunos de los factores que se suelen emplear.
2. Factores relativos al vehículo asegurado. Aquí se pueden considerar variables como el precio, la potencia, el modelo, el tipo de combustión o las ayudas a la conducción que tiene el vehículo. Estos factores están ganando paulatinamente más peso con respecto a otros, por la introducción de dispositivos en los vehículos que corrigen y ayudan en la conducción previniendo la ocurrencia de siniestros.
3. Otros factores asociados al ámbito territorial como la zona de circulación o el número de kilómetros a conducir.

La prima pura se suele descomponer en dos partes que se suelen analizar de forma separada en una gran parte de modelos. Esto es:

$$Prima\ pura = Frecuencia \times Severidad$$

La frecuencia es una medida de la tasa en la que un siniestro ocurre asociado a un intervalo temporal.

La severidad es una medida del coste medio de los siniestros para una cobertura concreta.

La clasificación de las distintas técnicas de tarificación del seguro de automóviles se puede ordenar, en base a dos grandes criterios no excluyentes:

1. Por un lado, el nivel de sofisticación de la técnica estadística empleada. Aquí se encuentran desde análisis descriptivos univariables, en el nivel más sencillo, hasta modelos híbridos de *Machine Learning* y GLMs en la parte más compleja.
2. Por otro lado, el ámbito de aplicación del proceso de modelización. Aquí se puede separar entre aquellas técnicas que se centran únicamente en la derivación del precio técnico, de aquellas otras que tratan de modelar también el comportamiento del cliente frente a la prima propuesta. Dentro de estas últimas han cobrado especial relevancia en los últimos años los modelos de optimización de primas.

Atendiendo al nivel de sofisticación de la técnica se tienen los modelos que se presentan en la siguiente figura:

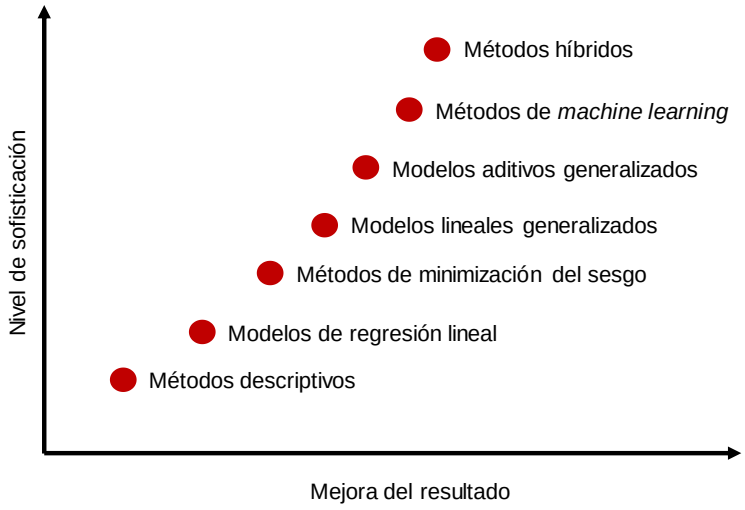


Figura 5.1: Nivel de sofisticación de modelos de tarifa

1. Métodos descriptivos

En esta clasificación entran aquellas metodologías que derivan el precio a través de la consideración de la descripción de la información histórica, normalmente para un número reducido de factores de riesgo. Se utiliza en mercados poco desarrollados, con bajo nivel de competencia, o en el que existen restricciones legales que prohíben la utilización de otras técnicas o la introducción de más variables.

La fijación de la prima indicada se puede realizar a través de la derivación directa del coste del riesgo por el cociente entre la siniestralidad y la exposición, a partir del coste y la frecuencia, o a partir de los ratios de siniestralidad observados.

La ventaja principal de estos métodos es su sencillez de cálculo, aunque presentan importantes inconvenientes como la deficiente segmentación del riesgo que provocan errores significativos de predicción, o en el caso de existan distintos factores, la no consideración del efecto de las correlaciones.

2. Modelos de regresión lineal

En esta clasificación se incorpora la derivación del precio a partir de la aplicación de técnicas de regresión lineal. Dentro de la categoría estarían los métodos de regresión lineal simple (en los casos de contar con un único factor) o múltiple si se cuenta con varias variables explicativas.

El inconveniente de la aplicación de esta técnica puede ser de nuevo la deficiente segmentación del riesgo y la consideración de la distribución del término de error como una variable aleatoria normal, que limita mucho su aplicabilidad en la fijación de la prima.

3. Métodos basados en la minimización del sesgo[12]

Estos métodos se iniciaron en los años 60 del pasado siglo. Consisten en el uso repetido de modelos univariables que tratan de, iterativamente, ir mejorando el resultado de la función de sesgo seleccionada hasta que ya no se obtiene mejora. Una explicación detallada de este método se puede encontrar en Feldblum y Brosius

[12] Conocidos en inglés como *Minimum Bias Procedure*.

(2002). En algunos mercados, se aplican derivaciones a esta fórmula de cálculo, conocidas también como métodos secuenciales, como ocurre en el caso del método de cálculo obligatorio de la prima en el seguro de automóviles en California.

Los métodos de minimización del sesgo no son técnicamente métodos multivariables, aunque la iteración sucesiva de esta metodología en muchos casos converge con los resultados obtenidos en modelos lineales generalizados como se demuestra en Mildenhall (1999).

4. Modelos lineales generalizados (GLM)

Los modelos lineales generalizados[13] se han convertido en la metodología estadística más extendida en la tarificación del seguro de automóviles. Está técnica se introdujo inicialmente en los años 70 del pasado siglo por Nelder y Wedderburn (1972) y los posteriores desarrollos que se recogen en McCullagh y Nelder (1989). La principal ventaja de estos métodos es que consideran el efecto de todas las variables simultáneamente e incluyen los modelos lineales clásicos como un caso específico, suavizando la restricción de normalidad de la variable respuesta, y dando la posibilidad que tome la forma de una distribución perteneciente a la familia exponencial. Adicionalmente, el efecto de los regresores sobre la variable respuesta no necesariamente debe ser aditivo si se transforma la escala.

5. Modelos aditivos generalizados (GAM)

En los seguros de automóviles hay distintos regresores, como la edad del conductor o el precio del vehículo que son variables continuas. En los GLMs las variables continuas se consideran como intervalos categóricos y los valores dentro de cada intervalo se consideran idénticos. Los modelos GAM, introducidos por Hastie y Tibshirani (1986), permiten trabajar con variables independientes numéricas sin tener que categorizarlas y mejorando la predictibilidad de las variables continuas (Kaivanipour, 2015).

[13] Además de los GLMs se han desarrollado extensiones de estos modelos como la de los modelos lineales generalizados mezclados (*generalized linear mixed models*) que se han usado para el conteo de datos binarios, *clusterizados* y datos longitudinales (Garrido y Zhou, 2009).

6. Modelos de *Machine Learning*

Se incorporan dentro de esta categoría los modelos de aprendizaje supervisado, aprendizaje no supervisado y los de aprendizaje reforzado. En esta categoría podrían considerarse entre otros los árboles de decisión, *Naive Bayes, K-nearest Neighbors, Support Vector Machine, Bagging y Ramdom Forest, GBM* o *XGBoost*.

7. Modelos híbridos

Se definen los modelos híbridos como aquellos que combinan técnicas tradicionales de tarificación con métodos de Inteligencia Artificial. Tal y como ya se refirió, estos modelos tratan de mejorar las predicciones que proporcionan las técnicas tradicionales, pero dotando a los resultados de mayor trazabilidad y transparencia.

Atendiendo al criterio del ámbito de aplicación de los modelos existen dos grandes divisiones. Por un lado, las técnicas actuariales clásicas para derivar la prima pura, por otro aquellas que modelan también el comportamiento del cliente frente a la prima propuesta.

Atendiendo al ámbito de aplicación de la modelización se tiene la clasificación que se presenta en la siguiente figura:

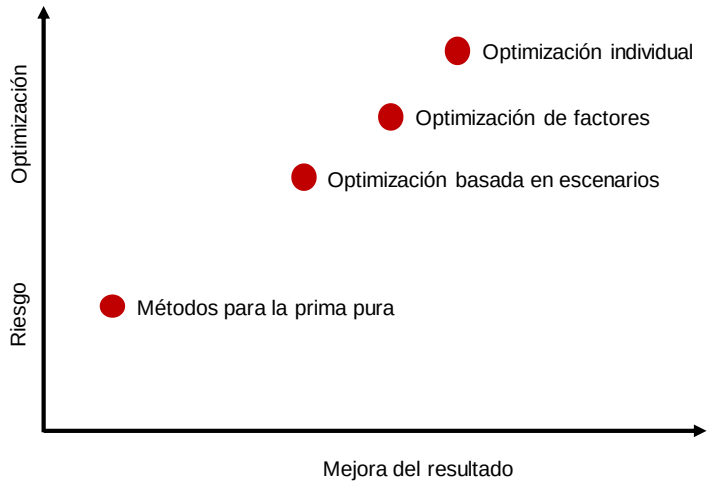


Figura 5.2: Ámbito de aplicación de los modelos de tarifa

En relación con los modelos de optimización existen las siguientes categorías:

1. Modelos de optimización basados de escenarios discretos

La introducción del comportamiento de los tomadores de seguro frente a la prima ofrecida, dentro del proceso de tarificación, lleva aparejado no sólo la predicción actuarial del riesgo, sino también la modelización estadística de ese comportamiento a partir de técnicas estadísticas. Estas técnicas se conocen en la jerga actuarial como la tarificación del comportamiento.

En la *Casualty Actuarial Society*[14], se define la optimización de precios como el complemento a los modelos tradicionales de riesgo incluyendo modelos cuantitativos de demanda para su uso en la fijación del precio de los clientes. La asociación americana de actuarios cataloga estos modelos como avanzados, y algunos estados de Estados Unidos[15] prohíben su utilización por considerar esta técnica discriminatoria y provocar variaciones en la prima no relacionadas con el coste del riesgo.

En definitiva, a los modelos que combinan ambos tipos de modelados, el actuarial y el del comportamiento, para derivar un valor óptimo de resultado o retención, sujeto a determinadas restricciones globales y parciales, se le denomina optimización de precios.

La aproximación más simple de los modelos de optimización consiste en calcular las primas en función de un set discreto de escenarios agregados para la cartera de pólizas. Estas estrategias se aplican a los modelos de riesgo y de retención[16] para cada uno de los escenarios. La estrategia seleccionada será aquella que más se acerque a los objetivos buscados por la empresa. Siendo estos modelos optimización, no son tan exactos por estar basados simulaciones finitas de escenarios. Además, se hacen sobre una cartera de pólizas de modo agregado.

[14] Véase (CASTF, 2015).

[15] Véase (ODI, 2015).

[16] Si se refieren a las renovaciones, o conversión para los de nueva producción.

2. Modelos de optimización de factores[17]

En estos algoritmos se trata de obtener la prima óptima según la función objetivo marcada por la entidad (típicamente maximizar el beneficio o la retención[18]). Para optimizar es necesario tener modelos de riesgo y modelos de retención, donde la elasticidad precio/demanda juega un papel muy relevante en la fijación del precio final.

En el caso de la optimización de factores se realizan los siguientes pasos:

1. Se realiza el proceso de optimización individual.
2. Con la estructura de factores y niveles de la tarifa se realiza un proceso de ingeniería inversa, donde se considera como variable dependiente la prima optimizada.

Tal y como se señala en Cummings (2015) esta técnica se aplica cuando la optimización individual de factores no es posible por restricciones regulatorias o por problemas en la implantación en los sistemas tecnológicos de las compañías. El principal inconveniente es que no es tan exacta como la optimización individual dado que el precio final se calcula por segmentos y no individualmente (Earnix, 2020).

3. Modelos de optimización individual

Los resultados de estos modelos se obtienen para cada póliza. Adicionalmente, tal y como se describe en Cummings (2015), presentan la ventaja de que no tienen por qué conservar la misma segmentación que la tarifa de primas sobre la que se está optimizando. Además, permite su aplicación en tiempo real, por lo que es la técnica que se suele utilizar para la nueva producción.

[17] Conocidos en inglés como *rating factor optimization*.

[18] En este apartado si en vez de optimizar las primas de cartera se optimiza la nueva producción en vez de modelos de retención se realizan modelos de conversión.

6. MODELOS ESTADÍSTICOS

En esta sección se presentan de forma concisa las técnicas y métodos más utilizados en la modelización predictiva de los seguros no vida.

En el proceso de modelización se pueden encontrar una serie de problemas. A continuación, se presentan los más comunes:

Sobreajuste u *Overfitting*

Ocurre cuando los modelos explican el ruido de los datos y no generalizan la relación intrínseca presente en estos. Se observa cuando el modelo explica de forma precisa los datos de entrenamiento, pero su poder predictivo es deficiente para el conjunto de datos de validación. Para esto, una de las cuestiones que se debe tener en cuenta es el balance entre sesgo y varianza, más conocido como *Bias-Variance Trade-off*. Este principio se basa en encontrar el punto óptimo que minimiza la relación entre ambos.

Adicionalmente, otro problema más sencillo de detectar es el infra ajuste, que no logra explicar el patrón en el conjunto de entrenamiento, lo que resulta en un error material en el conjunto de entrenamiento y de validación. El enfoque apropiado es generalizar el efecto de los datos de entrenamiento, con el objetivo de que sea capaz de explicar los datos que no se han usado. Estos problemas quedan muy claros en la siguiente infografía de la universidad de Stanford.

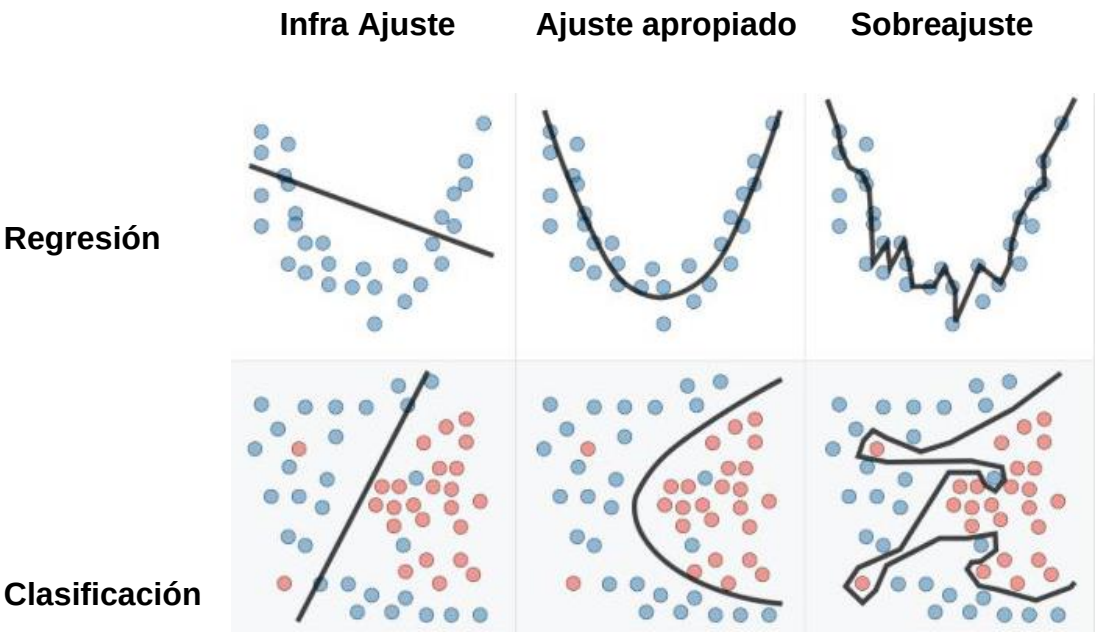


Figura 6.1. Ilustración de los tipos de error de modelo[19]

Adicionalmente, se distinguen dos tipos de errores:

- El error de entrenamiento, el cual disminuye a medida que aumenta la complejidad del modelo. Esto es un buen indicador de si el modelo está sobreajustado, ya que, si se alcanzan niveles de precisión superiores al 90%, pese a aparentar ser un gran resultado, es perjudicial porque los parámetros se terminan ajustando al ruido y características particulares de los datos.
- El error del conjunto de test permite detectar el problema mencionado en el punto anterior. Como regla general, a medida que aumenta la complejidad, debido a que el modelo se ha centrado en recoger el ruido del entrenamiento, tiende a tener menor precisión para datos no modelizados.

[19] Extraído de (Amidi y Amidi, 2018).

Con el objetivo de evitar estos problemas, en los modelos que se han realizado en este trabajo, se ha realizado una separación por muestreo aleatorio en dos submuestras, una de entrenamiento con 80% de los datos y otra de validación con el 20% restante.

Además, para las técnicas de *Machine Learning* se ha usado la técnica de *Cross-Validation*, que es un proceso iterativo que ejecuta la separación aleatoria un número determinado de veces y se utiliza para mejorar el resultado de los modelos.

Por lo tanto, se debe tener en cuenta que hay que conformarse con cierto poder predictivo, ya que superar ciertos umbrales suele implicar que el modelo está sobreajustado. Esta idea se conoce como *Early Stopping* y tiene como objetivo establecer unos parámetros que paren el entrenamiento del modelo para que no se vuelva demasiado complejo y se sobreajuste.

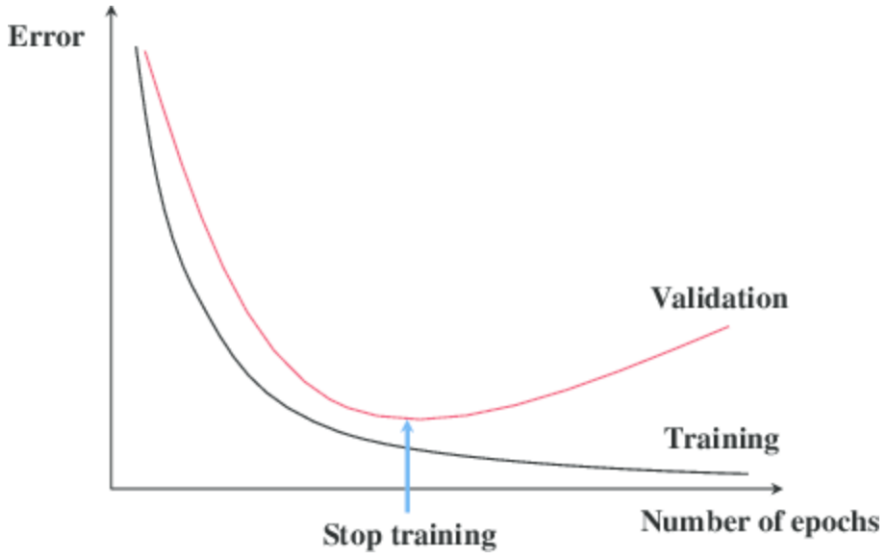


Figura 6.2. Ilustración del *Early Stopping*[20]

A continuación, se dividen los modelos en dos grandes grupos: modelos de regresión clásica y modelos de *Machine Learning*.

[20] Extraído de (Chen, 2020).

6.1 Modelos de regresión clásica

6.1.1 Modelos lineales generalizados (GLM)

Los modelos más utilizados por los actuarios de tarificación para explicar el comportamiento de las variables frecuencia y severidad siniestral son los modelos lineares generalizados o GLMs.

Su principal ventaja es la simplicidad e interpretabilidad. Además de esto, es robusto estadísticamente porque incluye técnicas como intervalos de confianza, contrastes de hipótesis entre otras.

El enfoque que se ha planteado es el uso de un modelo de frecuencia y severidad multiplicativo. Para estos modelos, es necesario tener una exposición superior a cero. Además, cabe destacar que para modelizar la severidad se deben eliminar aquellos registros donde la frecuencia es cero, esto conlleva la eliminación de una parte importante de la base de datos, debido a que la mayor parte de clientes no sufren siniestros, de lo contrario se obtienen valores infinitos.

En este trabajo, se han realizado cuatro modelos GLM. Uno de frecuencia para el *scoring* de vehículo, otro logístico para derivar la probabilidad de anulación por cliente y finalmente uno de frecuencia y otro de severidad, que se combinan para obtener la prima pura.

Se debe tener en cuenta que, al realizar los modelos siguiendo este enfoque, puede ser que las variables que son significativas para el modelo de frecuencia, como la potencia, no sean significativas en el modelo de severidad.

Adicionalmente, este tipo de modelos es fácil de implementar, ya que existen paquetes potentes en softwares libres como Python y R.

Los principales inconvenientes son que no incluye interacciones por defecto, así que depende del actuario que debe ser capaz de detectarlas e incluirlas. Además, requieren de experiencia y uso de juicio experto en su implementación, especialmente en la segmentación de las variables continuas.

En los modelos lineales clásicos, se asume que la variable objetivo sigue una distribución Normal con varianza constante. La ventaja de los GLMs es que se suavizan las asunciones del modelo, ya que permite que la variable respuesta pueda distribuirse según alguna de las distribuciones de la familia exponencial. Las asunciones que debe cumplir un GLM son las siguientes:

- Componente aleatorio. Los elementos de la variable objetivo "Y" son independientes y se distribuyen según uno de los miembros de la familia exponencial de distribuciones.
- Componente sistemático. Las variables explicativas combinadas permiten obtener el predictor lineal, tal que:

$$\eta = X \cdot \beta$$

- Función de enlace. Determina la relación entre las variables explicativas y la variable respuesta. Esta función es diferenciable y monótona, tal que:

$$E[Y] = \mu = g^{-1}(\eta)$$

Donde beta, mu y eta son vectores.

La función de enlace varía según el tipo de variable respuesta, a continuación, se muestran los tres tipos de modelos empleados en este trabajo:

Y	Modelo de frecuencia	Modelo de severidad	Modelo de renovación
Función enlace	ln(x)	ln(x)	ln(x/(1-x))
Error	Poisson	Gamma	Binomial

Tabla 6.1: Tabla con las funciones de enlace utilizadas

- Término *offset*. Este es un componente importante que se incluye cuando el efecto de una variable explicativa sobre la variable objetivo se conoce de antemano. Su derivación está realizada en la sección 7.6.4.

La no inclusión del componente *offset* es uno de los errores más comunes tal y como se describe en Tiwary (2020), "…y aunque parezcan similares, podemos concluir que usar el enfoque de modelización de la variable transformada sin incluir el componente *offset* es un enfoque erróneo".

A continuación, se deriva la estructura multiplicativa de los modelos GLM.

De forma general, los modelos lineales generalizados se definen tal que:

$$Y = g^{-1}(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_n \cdot X_n) + \epsilon$$

Dónde la variable dependiente toma la siguiente forma:

$$\pi_i = E(y = y_i | X = x_i)$$

Reorganizando la expresión algebraicamente, se usa una función enlace que transforma los regresores dependiendo de la variable respuesta, tal que:

$$g(\pi_i) = ln(\pi_i)$$

$$g\left(\pi_i\right) = \beta_0 + \sum_{j=1}^{n} \beta_j \cdot x_{ij}$$

Por lo tanto, la ecuación final de estructura multiplicativa es la siguiente:

$$\therefore \pi_i = e^{\sum_{i=1}^{n} \beta_0 + \beta_i x_i} = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\beta_2 x_2} ... \ e^{\beta_n x_n}$$

De esta forma, mediante la estructura multiplicativa, es fácil interpretar como afecta cada factor a la variable estudiada y facilita su implementación en los sistemas operacionales de suscripción de las entidades aseguradoras.

En el caso de los modelos de retención/conversión se utilizan GLM con función de enlace logit, donde la variable respuesta se distribuye de forma binomial. Su estructura es la siguiente:

$$f(z) = \frac{e^z}{1 + e^z}$$

Donde:

$$z = \beta_0 + x_1 \ \beta_1 + x_2 \ \beta_2 ...$$

Este modelo predice la probabilidad de fuga de los clientes, la derivación algebraica de la combinación lineal se presenta a continuación:

$$log\left(\frac{p}{1-p}\right) = z$$

Con el objetivo de resolver la incógnita p, se obtiene la siguiente expresión:

$$\left(\frac{p}{1-p}\right) = e^z$$

$$p = e^z \ (1-p)$$

$$p + p \ e^z = e^z$$

$$\therefore \quad p = \frac{e^z}{1 + e^z}$$

■

Para concluir, uno de los requisitos es que se tiene que cumplir la asunción de independencia entre variables explicativas, ya que de lo contrario estos modelos podrían tener problemas de multicolinealidad. Si no hay relación lineal entre los regresores, se dice que estos son ortogonales. Sin embargo, en la mayor parte de las aplicaciones de regresión los regresores no son ortogonales entre sí.

A veces no es grave la falta de ortogonalidad, pero en algunos casos, los regresores tienen una relación lineal casi perfecta y las inferencias basadas en el modelo de regresión pueden ser engañosas o erróneas.

6.1.2 Modelos aditivos generalizados (GAM)

Estos modelos van un paso más allá con respecto al modelo lineal generalizado. Con este enfoque se solventa el problema de tramificación de las variables numéricas que tienen los GLMs. Con respecto a su estructura, usan un enfoque semi-paramétrico tal que:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p_{cat}} \beta_j \cdot X_{ij} + \sum_{j=p_{cat}+1}^{p} f_{ij}(x_{ij})$$

Sin embargo, estos modelos pierden la interpretabilidad de los GLM y en ocasiones es difícil converger a un resultado computacionalmente.

Entre sus principales aplicaciones, estos modelos se usan para técnicas de zonificación. A continuación, se presenta el resultado de un caso práctico del Instituto de Actuarios Británico (IFoA) [21].

[21] Extraído de (Maréchal, 2020).

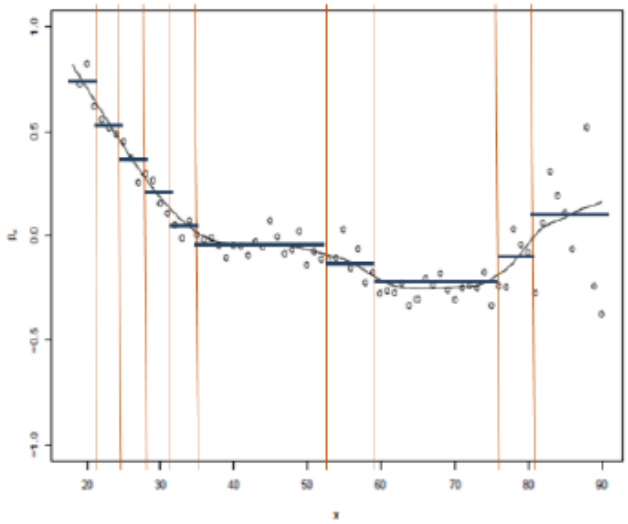


Figura 6.3. Segmentación de la frecuencia por zonas geográficas[22]

6.2 Modelos de *Machine Learning*

En los últimos años, en el ámbito actuarial, se han comenzado a utilizar algoritmos de una parte de la Inteligencia Artificial conocida como *Machine Learning*.

El objetivo final es “descubrir por sí mismos” el método que mejor prediga la información histórica propuesta.

Existen dos grandes grupos en cuanto a la variable objetivo a estudiar:

- **Clasificación:** Se dividen los resultados en dos o más grupos y el objetivo es clasificar observaciones no incluidas en el modelo en la categoría correcta. Algunos casos prácticos son: Detectar fraudes, predecir fugas en el momento de renovación entre otros.
- **Regresión:** Este enfoque se utiliza para las variables numéricas. Algunos ejemplos son el uso de modelos para predecir la frecuencia y la severidad de los siniestros entre otros.

[22] Extraído de (Maréchal, 2020).

6.2.1 Árboles de clasificación o regresión (CARTS)

Los CARTS funcionan de forma recursiva realizando particiones del conjunto de entrenamiento para conseguir subgrupos lo más puros posibles dada una variable objetivo.

Este algoritmo es la unidad básica para construir algoritmos más avanzados. Permite segmentar la variable respuesta en regiones homogéneas usando los distintos valores de las variables explicativas. Cada una de las regiones o nodos representan la media de las observaciones de cada clase. El problema de este algoritmo es que, por sí mismo, genera predicciones muy pobres y con gran variabilidad. Sin embargo, es un componente clave en algoritmos más avanzados de las familias de *Bagging* y *Boosting*.

El algoritmo selecciona las variables más discriminantes y las utiliza para separar los datos en grupos con cada vez menos variabilidad. Este proceso se ejecuta hasta alcanzar un nodo hoja, que son los puntos donde el algoritmo es incapaz de segmentar más los datos, o que se active uno de los criterios de *Early stopping*, que se establecen como cortafuegos para evitar que el modelo aprenda sobre el ruido del conjunto de datos del entrenamiento.

Otra técnica utilizada en este algoritmo es el podado, mejor conocido como *pruning*, cuyo objetivo es reducir el tamaño y la complejidad de los modelos para evitar precisamente el problema del sobreajuste.

A continuación, se explica el funcionamiento del algoritmo.

Se parte de f(x) como la función objetivo de la que se va a segmentar en grupos homogéneos.

$$f(x) = \sum_{m=1}^{M} c_m \, I(x \in R_m)$$

Donde c_m es un estimador de la media del grupo dada la partición.

$$\hat{c}_m = Media(y_i | x_i \in R_m)$$

Para cada variable “m” y de forma numérica se iteran las posibles particiones y se selecciona aquella variable que minimice el SSE. Los grupos creados segmentarán los conjuntos tal que:

$$R_1(j,s) = \{X | X_j \leq s\} \quad R_2(j,s) = \{X | X_j > s\}$$

La métrica que se utiliza es el error cuadrático medio y la fórmula iterativa minimiza el siguiente resultado:

$$min_{j,s} \left[min_{c_1} \sum_{x_i \in R_1(j,s))}^{x} (y_i - c_1)^2 + \ min_{c_2} \sum_{x_i \in R2(j,s))}^{x} (y_i - c_2)^2 \right]$$

Cabe destacar que $\hat{c}_1$ y $\hat{c}_2$ son los estimadores resultantes de realizar la media de los elementos pertenecientes a los dos subgrupos segmentados.

$$\hat{c}_1 = Media(y_i | x_i \in R_1(j,s)) \qquad \hat{c}_2 = Media(y_i | x_i \in R_2(j,s))$$

Además, este algoritmo casi siempre se realiza con segmentación en dos subgrupos, debido a que la complejidad computacional del problema es mucho menor cuando se bifurca en dos particiones.

Como este algoritmo siempre llega a los nodos más bajos del árbol generará *overfitting* si no se toman medidas. Para reducir el árbol y hacer el podado o *pruning* se utiliza el "podado coste-complejidad". El objetivo es que el árbol prediga bien en datos no usados para el conjunto del entrenamiento.

El árbol con todos los desarrollos posibles, sin ninguna medida de control, es T_0. Por ello, cualquier árbol resultante T es un subconjunto de T_0, tal que: $T \subset T_0$.

Siguiendo esta premisa, N_m es el número de observaciones que acaban en cada subgrupo. Esto permite obtener la media por subgrupo.

$$N_m = Number\left\{x_i \in R_m\right\}$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

Asimismo, existe un hiperparámetro que define el criterio para seguir realizando particiones conocido como coste de complejidad.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

El objetivo de este es, para cada alfa, obtener el subárbol que minimice $C_\alpha(T)$.

El ajuste del hiperparámetro alfa $\alpha \geq 0$ controla el coste de oportunidad entre bondad de ajuste y el tamaño del árbol. Cuanto mayor sea alfa, menor será el árbol y viceversa. Cuando alfa es cero no se penaliza nada y se obtiene el caso del árbol completo T_0 .

Para estimar alfa se ha usado el método de validación cruzada diez veces, que es el estándar, y se ha elegido el valor que minimice la suma de cuadrados.

Las principales ventajas de este algoritmo son que es muy visual y fácilmente interpretable. El tiempo computacional es relativamente rápido y es sencillo identificar las variables más significativas e interacciones en los datos. Adicionalmente, la presencia de datos nulos no supone un problema. Por otra parte, su principal desventaja es que presentan una gran varianza, lo cual puede conducir a malas predicciones.

6.2.2 Familia de algoritmos Bagging

Los métodos de ensamblaje, conocidos comúnmente como *Bootstrapping Agregating Methods* o simplemente *Bagging* es un enfoque propuesto por Breiman en 1996. La idea es simple, el algoritmo combina y ejecuta la media entre múltiples modelos de árbol usando la técnica de remuestreo *boostrap*. Esto reduce la variabilidad y el sobreajuste, lo que mejora considerablemente el poder predictivo. La idea principal es que, a mayor número de árboles, mayor robustez tendrá la predicción.

La metodología propuesta por Breiman para el caso de regresión es sencilla.

1) Se generan “m” muestras usando el remuestreo *boostrap* de un conjunto de datos de entrenamiento. Esto permite generar multitud de bases de datos similares, pero con variaciones aleatorias.

2) Para cada muestra, se ajusta un árbol CART sin limitaciones, que permite obtener los nodos hoja finales de cada subconjunto.

3) Finalmente, se calculan las predicciones individuales para cada árbol y se hace la media de este.

La diferencia con respecto a un caso de clasificación sería que en el tercer punto se usaría el criterio de la mayoría, es decir se escogería el valor más común entre todas las predicciones de los árboles individuales.

Uno de los algoritmos más usados de esta familia son los Bosques Aleatorios o *Random Forest*, que incorporan un cambio que hace que las predicciones mejoren considerablemente respecto al uso del *Bagging* clásico. La mejora consiste en que en el *Random Forest* no se consideran todas las variables para cada división, sino que se seleccionan un subconjunto del total de forma aleatoria cada vez, mientras que en el *Bagging* clásico siempre se consideran todas las variables y lo único aleatorio es la parte del muestreo *boostrap*.

El número de variables seleccionado para cada partición se determina de antemano. Una de las prácticas comunes es elegir el valor resultante de hacer la raíz cuadrada del número total de variables.

6.2.3 Familia de algoritmos Boosting

La familia de algoritmos *Boosting* combina una gran cantidad de modelos débiles para conseguir una gran capacidad predictiva. La idea es que se entrena estos modelos de forma secuencial, usando lo aprendido de sus predecesores para seguir mejorando.

La metodología que se ha usado en este trabajo es la *Gradient Boosting.* Lo particular de este tipo de modelos es que utiliza los residuos de los modelos anteriores para lograr una predicción más ajustada. En particular, se han usado varios algoritmos *Gradient Boosting Machine*, que están posicionados como unos de los más potentes en *Machine Learning* ya que con frecuencia resultan vencedores en competiciones de la plataforma *Kaggle*.

Para encontrar la solución óptima, se utiliza el algoritmo de descenso del gradiente. Esta técnica permite encontrar el mínimo global, que minimiza la función de pérdida tomando el error de forma recursiva y guiándose mediante el gradiente.

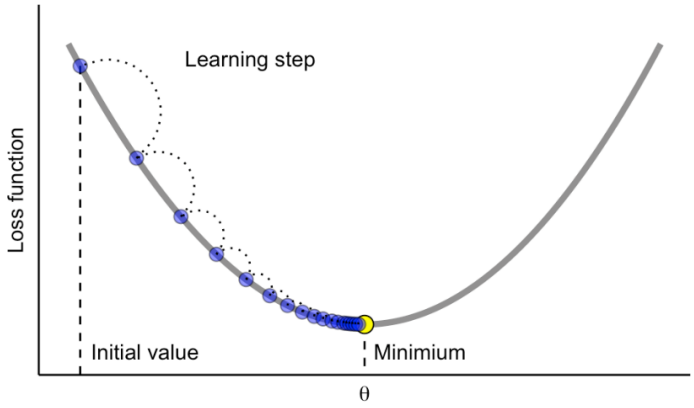


Figura 6.5. Ilustración del descenso del gradiente[23]

[23] Extraído de (Boehmke y Greenwell, 2020).

Sobre este punto, hay que destacar uno de los parámetros más importantes, la tasa de aprendizaje. Este determina el tamaño que tendrá ese paso que busque el mínimo local. Un valor muy pequeño hará el proceso mucho más lento y se requerirán muchas iteraciones para encontrar el mínimo. Por el contrario, si la tasa toma un valor alto podría saltarse el mínimo y acabar en una posición lejana a esta.

Las ventajas principales de estos métodos es que logran las predicciones más precisas, son flexibles ya que permiten optimizar funciones de pérdida diferentes y se pueden generar múltiples combinaciones de modelos según los hiperparámetros utilizados. En particular, los GBM no requieren de un preprocesado de datos, funciona tanto con variables numéricas como categóricas y no requiere de la eliminación o tratamiento de los registros *missing*.

No obstante, estos métodos minimizarán todo error que detecten, por lo que si no se ejecutan adecuadamente generarán un modelo sobreajustado. Por ello, se suele recomendar el uso de la validación cruzada. Otro problema es el tiempo de convergencia, debido al requerimiento computacional y a que cuando se calibran se comprueban muchas combinaciones de hiperparámetros simultáneamente.

Estos modelos tienen un problema de falta de transparencia y por lo tanto se les califica como “caja negra” por su dificultad a la hora de explicar el resultado.

6.3 Optimización matemática

Una de las cuestiones más interesantes es entender cómo se determina el precio o valor de los productos o servicios. Con el objetivo de entender este concepto, a continuación se explica en que consiste la optimización matemática de precios.

Grandes compañías como Booking, Uber o Amazon utilizan estas técnicas para maximizar el beneficio de sus compañías.

Sin embargo, antes es necesario entender uno de los conceptos principales de economía, la elasticidad precio demanda.

La demanda es una función monótona decreciente, a medida que el precio de un bien aumenta, su consumo disminuye. Por ello, la elasticidad se entiende como la sensibilidad de la demanda por cambios en el precio, tal que:

$$Elasticidad\ precio/demanda = \frac{Cambio\ relativo\ de\ demanda}{Cambio\ relativo\ en\ precio} = \frac{\frac{\Delta D}{D}}{\frac{\Delta P}{P}}$$

A continuación, se ilustra un caso práctico de cómo funciona la elasticidad al precio.

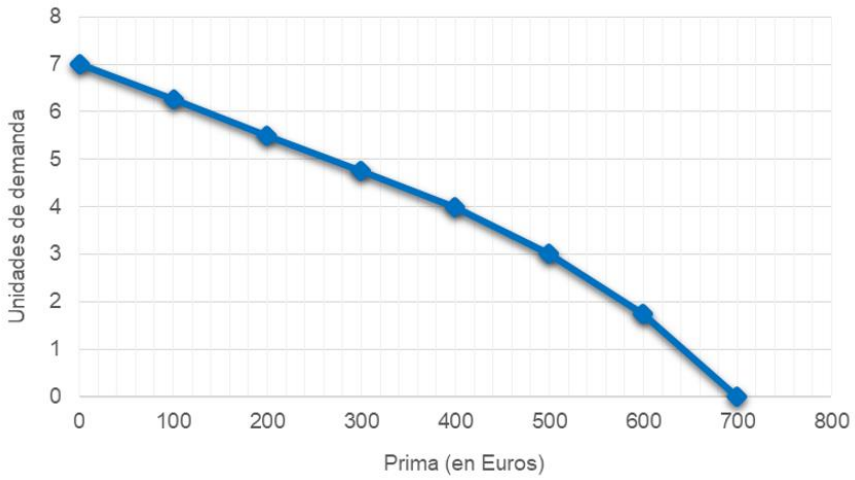


Figura 6.6. Variación de la demanda ante cambios de precio

Prima ofrecida	Unidades de demanda
0	7
100	6.25
200	5.5
300	4.75
400	4
500	3
600	1.75
700	0

Tabla 6.2. Evolución de la demanda con respecto al precio

En este ejemplo teórico, se aprecia como cuando se ofrece una prima de 300 euros, por cada 1% de aumento de la prima la demanda cae un 0,47%. Asimismo, cuando se ofrece una prima de 500 euros, cada aumento del 1% sobre el precio supone una bajada de la demanda de 2%.

La optimización matemática consiste en la maximización de una variable objetivo mediante técnicas numéricas siguiendo las pautas establecidas por la dirección de la entidad.

En la industria aseguradora, para optimizar primas es necesario disponer de la siguiente información:

- Modelos de coste, que predigan la prima pura y otros costes para los distintos perfiles de cliente.

- Análisis competitivo, que compare el posicionamiento de la compañía *versus* el mercado.

- Modelos de la elasticidad de los clientes, que deben tener en cuenta los precios de la competencia y el comportamiento del cliente dependiendo sus características individuales y la situación de la industria.

La optimización de precios es ampliamente utilizada en el seguro de automóviles, ya que este es un mercado muy maduro y competitivo, y hay que utilizar soluciones innovadoras que mejoren el resultado, aunque únicamente suponga una mejora marginal con respecto a los competidores.

Es una práctica muy utilizada en compañías grandes con volumen suficiente y se lleva poniendo en práctica desde principios de este siglo (Santoni y Gómez, 2007).

Mediante la combinación de modelos de coste y elasticidad para los distintos canales de distribución y segmento de cliente se puede maximizar la relación entre beneficio por póliza y volumen requerido por la compañía para alcanzar los KPIs marcados por la dirección.

Se puede resumir el proceso de optimización en los siguientes puntos:

1) Alinear los objetivos de restricciones con la dirección de la empresa.
   Existen distintas estrategias, resumidas principalmente en maximización del margen esperado o la maximización de la retención de la cartera.

   Adicionalmente, se pueden establecer restricciones de tipo global y/o individual. Algunos ejemplos de restricciones son:

   - Restricciones globales, como el mantenimiento de la cartera en vigor con una tasa de retención del 70% o disponer de un margen medio por póliza de 60 euros.

   - Restricciones individuales, como limitar las variaciones de precio respecto a la prima previa no pueden sobrepasar un rango determinado, por ejemplo [-10%,+10%].

2) Gap análisis
   Se emplea el modelo de *pricing* de la entidad como input para el proceso de optimización de la compañía. Valora las diferencias entre las primas no optimizadas y las optimizadas usando tests A/B.

3) Benchmarking con la competencia.
   El posicionamiento de la entidad en el mercado supone una variable fundamental en la optimización. Este estudio de mercado permite analizar en que segmentos la compañía es más barata o cara con respecto a sus competidores, o como de competitivo es por segmento. El precio de los competidores se utiliza como una variable explicativa en los modelos de retención/conversión.

4) Análisis de las renovaciones.

   La información histórica de renovaciones se usa para determinar el comportamiento de los clientes en el modelo de retención. Una aplicación práctica se ha realizado en el apartado 7.7 de este trabajo.

   Para el modelo de demanda, se pueden utilizar variables que se pueden segregar en varios grupos:

   - Variables de la póliza: Como diferencia relativa entre el precio actual y el propuesto en la renovación, tipo de producto y coberturas, diferencia relativa entre la prima ofrecida y la media de la ofrecida por los principales competidores, número de años en la compañía y canal entre otras.
   - Variables del perfil de riesgo del asegurado: años sin siniestros, edad del asegurado, género[24], años de experiencia como conductor, entre otras.
   - Otras: Que incluyen *scoring* crediticio, frecuencia de pago, antigüedad en la compañía entre otras.

5) Optimización.

   Existen distintas de técnicas para realizar la optimización de primas. Para funciones no lineales algunos softwares comerciales utilizan algoritmos numéricos sofisticados o el método de derivadas parciales usando el Lagrangiano.

   En este trabajo, se ha utilizado el enfoque de optimización individual usando el algoritmo *Grid Search*, que se explicará en el caso práctico.

[24] Pese a que el Tribunal de Justicia Europeo dictaminó en 2012 que está prohibido diferenciar la prima ofrecida por sexo, en algunas jurisdicciones en el mundo se permite el uso de esta variable, que resulta ser significativa para el seguro de automóviles.

6) Implementación de la estrategia de tarificación de la compañía.
   Esta técnica permite encontrar el valor óptimo de prima para cada cliente en función de la métrica seleccionada por la dirección.

7. UNA APLICACIÓN EMPÍRICA

7.1 Introducción

En este trabajo se propone un proceso de modelización de optimización individual de primas del seguro de automóvil de la modalidad de seguro Todo Riesgo. Esto implica la realización de distintos modelos predictivos. Tal y como se recoge en Santoni y Gómez (2007) y en Spedicato et al. (2018), para los procesos de optimización es necesario realizar modelos de coste para predecir el importe esperado de la siniestralidad por cliente, modelos de competencia, modelos de demanda y por último la agregación de todos los algoritmos a partir de la técnica de optimización elegida.

Más concretamente en este estudio se van a realizar seis modelos predictivos diferentes que se agregan finalmente en un algoritmo de optimización. Estos son:

1. Modelos de *scoring* de vehículo. En este punto se enfrentan dos algoritmos estadísticos diferentes para medir su naturaleza predictiva. Por un lado, se realiza un *scoring* a partir de un GLM de frecuencia. Por el otro se compara el modelo resultante con un GBM y se selecciona el más predictivo.

2. El anterior modelo de clasificación de vehículos se introduce como variable independiente en los modelos GLM de frecuencia y de coste medio obteniéndose el modelo de prima pura.

3. Utilizando la técnica de GBM se derivan los precios de los competidores. Con carácter previo se ha elaborado una base de datos con precios de la competencia a partir del empleo de técnicas de *web-scraping*.

4. Con los precios derivados a partir del GBM y con otras variables adicionales sobre las renovaciones de los clientes se obtienen las curvas de demanda para cada cliente a través de un GLM logístico.

5. Integrando la información del comportamiento del cliente y del entorno de mercado con los modelos tradicionales de siniestralidad se obtiene la prima

óptima a partir de la técnica de optimización individual. La metodología utilizada emplea un conjunto de escenarios discretos como se propone en Spedicato et al. (2018) con la técnica de *Grid Search optimization*[25].

Expresando todo el modelado predictivo de optimización de modo gráfico se tiene:

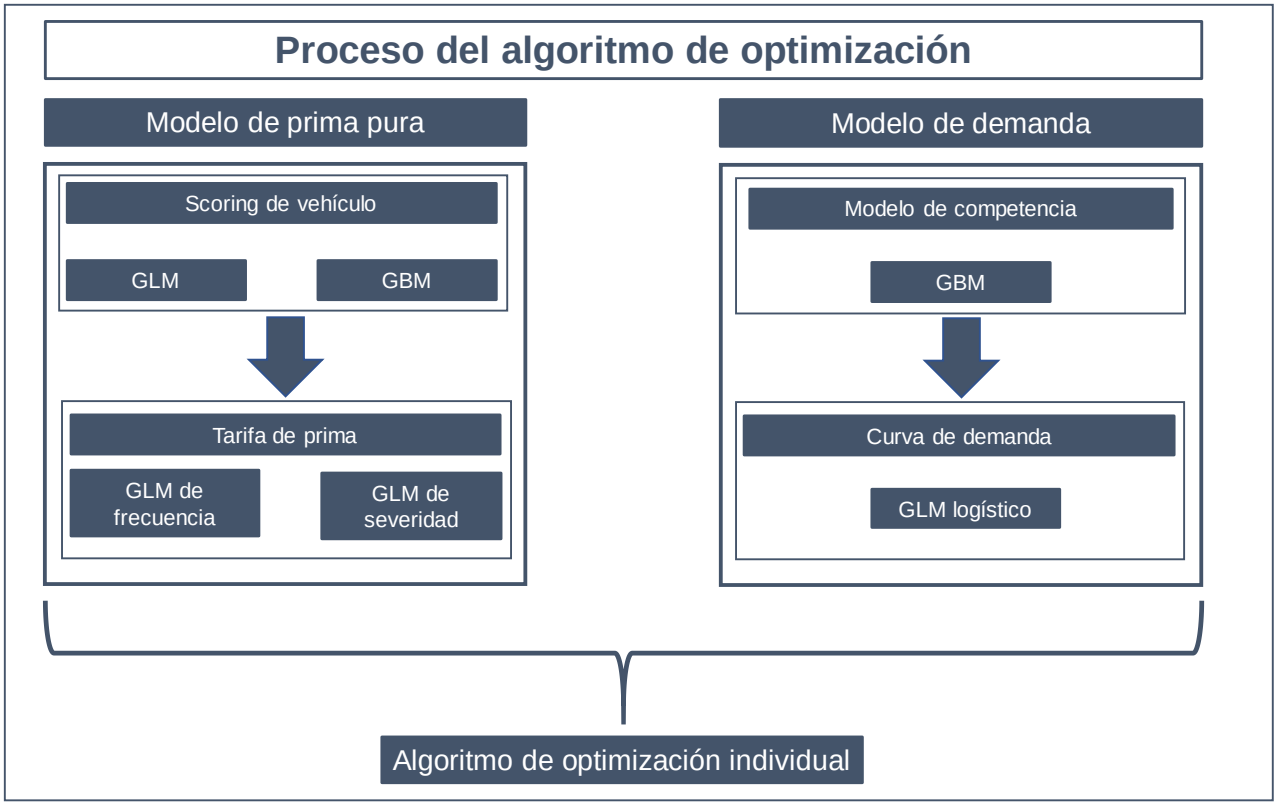


Figura 7.1: Esquema de los modelos predictivos del proceso de optimización

7.2 Software utilizado

La práctica actuarial, y en concreto los actuarios de tarificación de las entidades aseguradoras, utilizan muy frecuentemente software comercial para la derivación

[25] En Earnix (2020) se le denomina así a esta técnica por lo que se usa esta expresión en inglés para denominar este tipo de optimización.

de las primas. Como elementos positivos de estas herramientas de software[26] están que permiten el empleo de distintas técnicas de un modo sencillo y añaden adicionalmente muchas funcionalidades que ayudan en el proceso de elaboración de los modelos. En el lado negativo, estos softwares propietarios presentan elementos poco explicables, que los pueden convertir en “cajas negras”, y sus licencias tienen un coste que puede resultar prohibitivo para algunas entidades. En los últimos años, comienza a ser frecuente la combinación de estos softwares comerciales con softwares libres como R o Python.

Uno de los objetivos de este trabajo consiste en realizar un proceso completo de modelización avanzada de tarificación en el seguro de automóviles con el uso exclusivo de software libre. Los programas utilizados en este trabajo han sido los siguientes:

1. El software libre R para la limpieza de datos y modelización. Los paquetes utilizados se han clasificado en varios subgrupos principales:
    - Tidyverse. Esta colección de paquetes introducidos por Hardley Wickham y su equipo incluyen una gran selección de librerías que mejoran y facilitan la transformación, modelización y visualización de datos.
    - Cluster y Rattle. Estos paquetes se han utilizado en la sección de aprendizaje no supervisado del punto 7.5.1.
    - GBM, H20, Caret y Rpart. Estos paquetes se han usado en la modelización de los algoritmos de *Machine Learning* que se han empleado.
    - Para la modelización de los GLM, se han usado una serie de paquetes implementados por Chamber y Hastie en 2002 en el paquete Stats. También se han utilizado las funciones del paquete MASS, creado por Venables y Ripley en 2002.

[26] Aunque hay distintas firmas comercializando software actuarial el más utilizado en el mercado es Emblem y Radar, de Willis Towers Watson, y Earnix.

- Se han utilizado los paquetes de Shiny y Shinydashboards que permiten la creación de herramientas reactivas online para interpretar los resultados de los modelos de manera muy visual.

2. El software libre Python para realizar la *web-scraping* permite derivar la tarifa de la competencia sin acudir a proveedores externos de precios, en particular el paquete utilizado ha sido el entorno de pruebas Selenium.

3. Se ha empleado Microsoft Excel para la creación de visualizaciones específicas y excepcionalmente para el tratamiento de los datos.

7.3 Los datos

La información utilizada en este trabajo proviene de una entidad aseguradora que opera en el negocio de automóviles en España. La base de datos disponible se corresponde con la información de renovación de los años de calendario 2018 y 2019, para vehículos de primera categoría comprendiendo un total de 155.157 registros de dos productos diferentes de la modalidad de Todo Riesgo, uno con franquicia y otro sin franquicia. La información proporcionada se refiere al uso particular de los vehículos y consta de distintas bases de datos:

1. Base de datos con información sobre las características sobre la póliza, el tomador y el conductor del vehículo. Se incluyen variables como antigüedad de la póliza, coberturas contratadas, edad del conductor, género, antigüedad del carnet, historial crediticio del tomador, zona de circulación o el nivel de bonus malus.

2. Base de datos con información sobre las características técnicas de los vehículos. Incluyen variables como años del vehículo, peso, Potencia, marca, modelo o precio.

3. Base de datos con información sobre número e importe de siniestralidad por tipo de siniestro. Se modelará únicamente la cobertura de daños propios.

4. Base de datos con información de los precios de los competidores.

La información se ha unificado mediante el identificador de póliza. Adicionalmente, cabe destacar que los datos han sido anonimizados con el objetivo de cumplir con la ley de protección de datos.

Adicionalmente, se asigna el 80% de los datos al entrenamiento del modelo y el 20% restante se utilizará que se harán las validaciones de los modelos.

7.3.1 Enriquecimiento con datos externos

Los datos han sido enriquecidos con dos fuentes externas de datos. Estas fuentes exógenas a la compañía añaden información útil, que contribuye a una mejora de la precisión de las predicciones. Esto constituye una práctica muy habitual de las compañías de seguros que utilizan estas fuentes externas para recabar información adicional de los asegurados, que supone una ventaja competitiva en un ramo tan competitivo como el de automóviles.

La primera es la variable de *credit scoring* suministrada por un proveedor externo. Es una de las variables más importantes en la modelización porque recoge información que las compañías no recogen en sus sistemas. Es una variable que está poco correlada con la información con la que cuenta la compañía y esto aporta una capacidad predictiva adicional, especialmente relevante en la evaluación del fraude y la siniestralidad (Miller y Smith, 2003).

Adicionalmente, se han obtenido precios de competidores en internet a través técnicas de *web-scraping*[27] utilizando el entorno de pruebas software libre Selenium. Estos datos han servido para alimentar un modelo de predicción de las tarifas de la competencia para considerar esta variable en los modelos de demanda de los tomadores. En concreto, se ha configurado un *bot* en Python que ha extraído

[27] En este trabajo se ha empleado esta técnica para obtener precios de distintos competidores. La técnica de *web-scraping* puede emplearse para obtener múltiples tipos de datos que pueden ser buenos predictores dependiendo del problema de modelización que se quiera resolver.

100.000 registros base de un *model point* y se han usado las tres primeras cotizaciones de cada uno para sacar una media del mercado. Después de esto, se ha usado un modelo de GBM para inferir la tarifa creada a partir de la media de las primas de la competencia.

7.3.2 Preprocesamiento de los datos

Esta fase de la modelización consiste en la transformación de determinadas variables, como el cálculo de la exposición, así como la depuración de los datos.

Todas las variables categóricas han sido transformadas a factores, esto se debe a que una parte de los algoritmos de *Machine Learning* que se van a utilizar requieren que todas las variables no numéricas tengan la estructura de factores.

Asimismo, para tratar con el software se han dividido en filas considerando años naturales. Esto es, una póliza que empezase en Julio del 2018 tendría una exposición de 6 meses en 2018 y los otros 6 meses de exposición serían atribuibles a 2019. Cabe recalcar lo que se entiende por exposición, que es el peso que se le asigna a una póliza mientras está en vigor. La exposición es una fracción que toma valores de entre [0,1], tal que, si se considera distribución uniforme de la siniestralidad, se tiene que:

$$Exposicion = \frac{Dias\ de\ la\ poliza\ en\ vigor\ durante\ el\ año\ natural}{365\ dias}$$

En ejemplo anterior, supondría que la póliza tendría un valor de 0,5 en 2018 y de 0,5 en 2019.

Asimismo, debido al tamaño de la base de datos se han distinguido varios subconjuntos a tratar según el modelo para el que se utilizarán:

1) Datos para los modelos GLM de frecuencia y severidad, para derivar la prima pura.

2) Datos para el modelo de demanda. Son aquellos mediante los cuales se puede derivar, la probabilidad de retención del cliente y la elasticidad al precio de este.

La mayor parte de estas variables se comprenden en el periodo de tiempo cercano a la renovación, por lo que la mayoría se utiliza únicamente para el modelo de demanda. Pese a esto, hay algunas variables muy importantes como la edad del tomador que se usan también en el modelo de riesgo.

3) Datos utilizados para constituir el *Car Group*. Se dispone de 25 variables relacionadas al vehículo. En este trabajo se determinará como influyen las características individuales del vehículo en la frecuencia siniestral de la cobertura de daños propios. Estas variables se han usado para crear dos modelos que se enfrentarán para comprobar su capacidad predictiva. Por un lado, un modelo GLM de frecuencia sobre el que se obtiene un *scoring*. Adicionalmente un modelo GBM con el que se realiza otro *scoring*. Ambos arrojan una puntuación que clasifica el vehículo según su riesgo (modelo de frecuencia). Se elige finalmente el modelo que clasifica mejor.

7.3.3 Depuración de los datos

Para la realización de los modelos de riesgo se ha filtrado la base de datos asegurando la coherencia en los mismos y se han buscado valores atípicos, que distorsionarían las predicciones de los modelos. Siguiendo los criterios que se proponen en Anderson et al. (2007) algunos de los cambios han sido:

- Exclusión de todas las pólizas que no tengan al menos un día de exposición.
- Comprobación de que las fechas de nacimiento son las mismas para los distintos momentos de la póliza a lo largo del tiempo, así como que las edades de los asegurados cambian de un año para otro con el paso de los años.
- Eliminación de siniestros de cuantías negativas o iguales a cero. Esto se debe a que este hecho no tiene sentido económico y adicionalmente la distribución de la severidad con una distribución Gamma no admite cuantías negativas.
- Exclusión de aquellas pólizas que no cumpliesen la siguiente desigualdad:

$$Edad - Antiguedad\ del\ carnet \geq 18\ años$$

- Eliminación de registros con primas menores o iguales a cero.

En los casos de valores nulos, estos se han rellenado con la mediana de esa variable en los casos de variables numéricas, mientras que se ha usado la moda para completar los registros *missing* en variables categóricas.

7.4 Metodología de modelización

En este estudio empírico se ha utilizado la estrategia de modelización que se propone en (Werner, 2005). Este procedimiento es el que se ha seguido en todos los modelos del tipo GLM realizados en este trabajo. El mecanismo de depuración sigue los siguientes puntos:

1. Depuración de los datos.
2. Selección de la estructura inicial del error y de la función de enlace.
3. Análisis exploratorio inicial.
4. Construcción recursiva de los modelos.
5. Validación del modelo final.

En los siguientes epígrafes se irán trasladando algunos de los resultados de los modelos finales. No obstante, dado el elevado número de modelos a realizar se ha optado por no adjuntar las salidas de todos los modelos[28].

[28] En los anexos adjuntos se presentan las salidas univariables de los distintos modelos.

7.5 Modelo de *scoring* de vehículo

Tal y como se trasladó en la introducción de este capítulo en la modelización primero se han realizado dos algoritmos de *scoring* sobre los datos de frecuencia de daños propios, uno con un GLM y otro con un GBM. Finalmente se ha seleccionado el modelo GBM que proporciona una clasificación más exacta.

Con la clasificación del vehículo y el resto de regresores seleccionados se han realizado unos modelos GLM de frecuencia y severidad creando un modelo híbrido de técnicas de *Machine Learning* y de modelos lineales generalizados.

7.5.1 Creación del scoring de vehículo a partir de GLMs

La continua evolución de la tecnología embarcada en los vehículos con las ayudas a la conducción, así como la introducción de otras características como el incremento del parque de automóviles eléctricos o la conectividad a internet, está provocando cambios en el efecto que tienen estas variables específicas del vehículo en la siniestralidad. Las variables tradicionales relacionadas con el conductor tenderán sin duda a perder peso en el futuro frente a aquellas que se relacionan con las ayudas a la conducción de los vehículos. En ICEA (2018) se trasladan algunas cifras y se indica como, en el nivel más maduro de coche autónomo, la siniestralidad podría llegar a reducirse en un 95%.

En este entorno cambiante tiene sentido incorporar en los modelos una variable que recoja el efecto de las distintas características del vehículo. El *scoring* de vehículo, o *Car Group*, es una variable sintética que se calcula a partir de las características individuales de cada vehículo. Esta metodología presenta varias ventajas frente a la utilización de las variables específicas del vehículo dentro de los modelos. Por un lado, permite una reducción de dimensionalidad que repercute en una mejora respecto a los criterios de selección de modelos, que penalizan la introducción de variables, como es el caso del criterio de información de Akaike (AIC) o el criterio de información Bayesiano (BIC). Adicionalmente, y como traslada Boison (2011), permite una tarificación más sencilla de los nuevos vehículos, permitiendo incorporar inmediatamente los cambios en las mejoras en las ayudas a la

conducción. Por tanto, pese a una cierta pérdida de interpretabilidad, estos métodos permiten una integración rápida de nuevas variables en los sistemas de emisión de las compañías aseguradoras.

En esta sección se han agrupado todas las variables significativas del vehículo, 12 variables, para crear el *scoring,* usando las características individuales de cada vehículo. De las doce variables, siete son variables continuas y cinco son categóricas.

Las variables utilizadas en el proceso de modelización han sido las siguientes:

- CARAGE: Representa la antigüedad del vehículo, es una variable numérica que toma valores dentro del rango [0,22].

- CARCC: Representa la cilindrada del vehículo, es una variable numérica que toma valores dentro del rango [0,5439].

- CARCV: Representa los caballos de vapor correspondientes al vehículo, es una variable numérica que toma valores dentro del rango [0,367].

- CARDOORS: Representa el número de puertas del vehículo, es una variable numérica que toma valores dentro del rango [0,6].
- CARFUEL: Representa el tipo de combustible que utiliza el vehículo, es una variable categórica en la que hay tres niveles diferenciados: Gasolina, Diésel y Eléctricos.

- CARGPS: Representa la presencia de GPS integrado en el vehículo, es una variable dicotómica con los niveles Sí/No.

- CARKW: Representa el número de kilowatios por vehículo, es una variable numérica que toma valores dentro del rango [4,270].

- CARMAKE: Representa la marca del vehículo, es una variable categórica en la que se tienen 46 marcas distintas.

- CARMODEL: Representa el modelo del vehículo, es una variable categórica en la que se observan 586 modelos distintos.

- CARPVP: Representa el precio del vehículo, es una variable numérica que toma valores dentro del rango [5812,119400].

- CARSEGM: Representa el segmento al que se asigna el vehículo. Es una variable categórica.

- CARTAR: Representa la tara del vehículo, es una variable numérica que toma valores dentro del rango [600,2669].

7.5.1.1 Selección de la función de enlace y la estructura de error

Para el scoring de agrupación de vehículo se ha considerado un modelo con una función de enlace log con una estructura de error Poisson.

Cabe destacar que la metodología de estimación de los coeficientes de todos los modelos se realiza por el método de máxima verosimilitud. Esta técnica de derivación de estimadores calibra los parámetros tal que, asumida la forma del modelo, producirá los resultados observados de la variable dependiente con la máxima probabilidad. Se entiende como el producto de las probabilidades de obtener los valores observados de la variable objetivo.

7.5.1.2 Análisis exploratorio inicial

En este apartado se presentan de modo univariable los principales regresores candidatos con sus niveles, en relación a la variable respuesta. Este análisis sirve para tener una idea preliminar de las variables candidatas a entrar en los modelos.

1. Tipo de combustión del vehículo

A continuación, se explora descriptivamente una de las variables más interesantes del trabajo, por la evolución futura que tendrá el desarrollo de los vehículos eléctricos.

En la actualidad, los datos observados[29] muestran una exposición baja para los vehículos eléctricos por lo que el nivel no debería considerarse en la modelización. No obstante, se observa que la frecuencia en la siniestralidad de daños propios de los vehículos eléctricos, que incluye eléctricos e híbridos, es menor que la de combustibles tradicionales de combustibles fósiles. Esta evidencia empírica de menos accidentes de tráfico ya se ha puesto de manifiesto en distintos trabajos como en (García, 2021).

Debido a la baja exposición de la variable, los errores estándar de la relatividad correspondiente a los vehículos eléctricos es alta. Sin embargo, las políticas europeas en materia de sostenibilidad medioambiental y las tendencias hacia vehículos de fuentes energéticas más sostenibles harán de ésta una variable a tener en cuenta. En efecto, en los próximos años se espera que estos vehículos representen un importante porcentaje del parque de vehículos vendidos, por lo que se incorporará en el *scoring* de vehículo, aunque el nivel de la exposición sea muy bajo.

[29] Los datos observados en la base de datos son compatibles con los que se presentan en UNESPA (2019) para todo el mercado español donde sólo un 1,8% del parque total de vehículos utiliza combustible no fósil (híbridos y eléctricos).

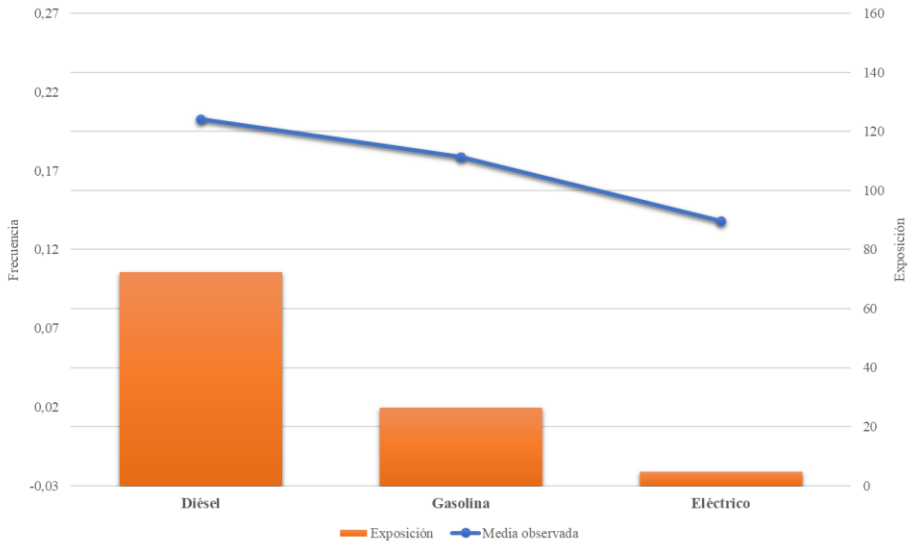


Figura 7.2: Distribución de la frecuencia según el tipo de combustible

2. Marca del vehículo

El gráfico descriptivo que se muestra para la frecuencia presenta un gran número de marcas de vehículo, que consecuentemente generan un sustancial número de niveles con poca exposición, tal y como se observa a continuación:

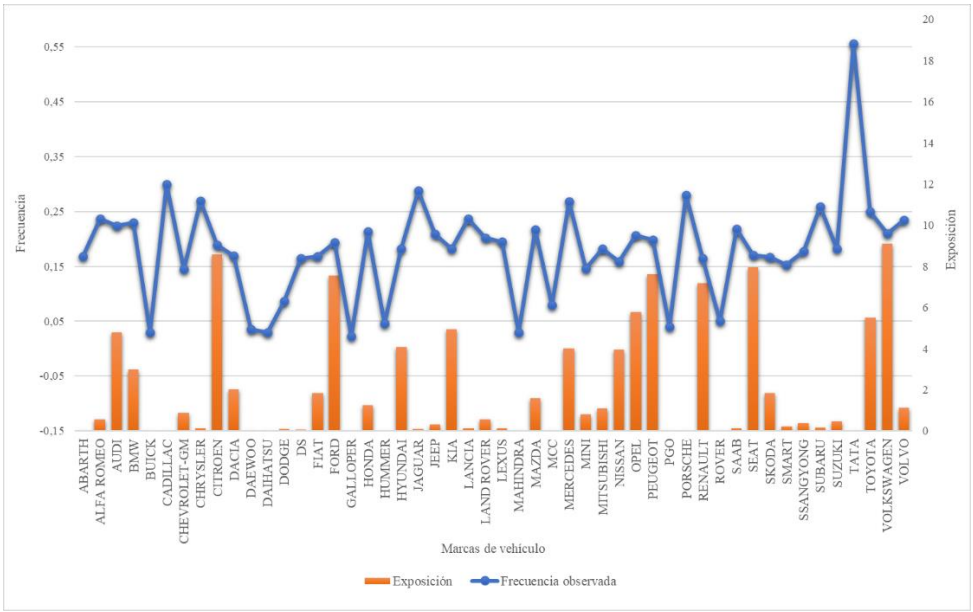


Figura 7.3: Distribución de la frecuencia según la marca del vehículo

Es importante destacar que ciertas variables como la marca del vehículo requieren un procedimiento para agrupar sus niveles. Para ello, se han utilizado dos metodologías, una técnica de aprendizaje no supervisado conocida como *Clustering* y una metodología más innovadora usando *Machine Learning*.

Es necesario agrupar los niveles en grupos, dado que, si se estimase un coeficiente para cada marca, la estimación que se haría sería muy inestable por el insuficiente volumen de datos para algunos segmentos.

En primer lugar, se presenta la agrupación usando *Clustering*. Tradicionalmente, en cálculo multivariante se utiliza la distancia Euclídea. El concepto estadístico de distancia es una medida que permite medir la proximidad de los individuos entre sí.

No obstante, la variable en cuestión, "CARMAKE", es categórica y la distancia Euclídea es únicamente aplicable a las variables numéricas. Por lo tanto, es necesario utilizar una métrica más avanzada que funcione en datos mixtos[30].

Debido a esto, el algoritmo propuesto es el *K-Medoids*, que es una técnica de agrupamiento emplea el concepto de distancia de Gower que minimiza la distancia de los puntos a una serie de *K* centroides determinados de antemano. Esta técnica fue introducida por primera vez en 1971 por J.C. Gower.

La distancia de Gower requiere de un concepto estadístico como es la disimilitud. Tras realizar esta estimación, se emplea la técnica del coeficiente de Silhoutte para determinar el número óptimo de clústeres que se usarán en la modelización.

La distancia de *Gower* se calcula como la media de las disimilitudes parciales entre distintos individuos, cada disimilitud parcial toma valores dentro del rango [0,1]. El cálculo depende de la variable estimada y requiere la normalización de cada variable, hay que reescalar los valores usando la fórmula de tipificación de cálculo multivariable:

[30] Hace referencia tanto a datos numéricos, como a datos categóricos.

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Y la distancia de *Gower* se representa tal que:

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$$

Para las variables categóricas, la disimilitud parcial es igual a uno si los valores son distintos y es cero sí coinciden.

El algoritmo de *K-Medoids*, similar al *K-Medias*, agrupa las observaciones en grupos homogéneos y permite elegir el número de centroides.

Asimismo, la técnica de t-SNE *(t-Distributed Stochastic Neighbor Embedding*) permite reducir la dimensionalidad para la visualización de bases de datos grandes.

Finalmente, se han identificado 3 centroides diferenciados que permiten explicar los modelos y las marcas de coche.

En segundo lugar y debido a la gran variabilidad dentro de las marcas de los vehículos, se ha propuesto utilizar el método más conocido en cuanto a árboles de regresión y se denomina CART o Árboles de Clasificación y Regresión (Breiman et al.,1984).

Estos métodos funcionan segmentando la base de datos en subconjuntos más pequeños e incrementando progresivamente su homogeneidad en cada paso. Individualmente, estos modelos son bastante inestables y son malos predictores, pero son muy visuales y sencillos de interpretar. Por este motivo, estas unidades básicas se combinan en técnicas más avanzadas agrupadas en las familias de algoritmos de *Bagging* y *Boosting*, que logran gran capacidad predictiva. La mayor parte de modelos avanzados que se usan en la actualidad pertenecen a estas dos familias.

En cuanto a los CART, es un algoritmo sencillo y fácil de entender, además es visual en cuanto a que se puede apreciar las particiones. El criterio de partición consiste en minimizar el error entre las dos muestras, tal y como se ilustra a continuación:

$$Suma\ de\ errores\ al\ cuadrado = MIN\left\{\sum_{i\in S_1}(y_i-\bar{y_1})^2+\sum_{i\in S_2}(y_i-\bar{y_2})^2\right\}$$

Es importante destacar que las particiones siguen una jerarquía, es decir dependen directamente de las segmentaciones que se han realizado anteriormente. Además, el criterio que determina la variable a utilizar se estima a partir de minimizar la suma de errores al cuadrado, se prueban todas las variables y se selecciona la más discriminante que reduzca la suma de errores al cuadrado. El algoritmo es bastante flexible en cuanto a la imposición de limitaciones y a como se determina el *Early stopping*, que previene el problema de sobreajuste[31].

[31] Una explicación de cómo funciona este método se puede encontrar en el punto 6 de este trabajo.

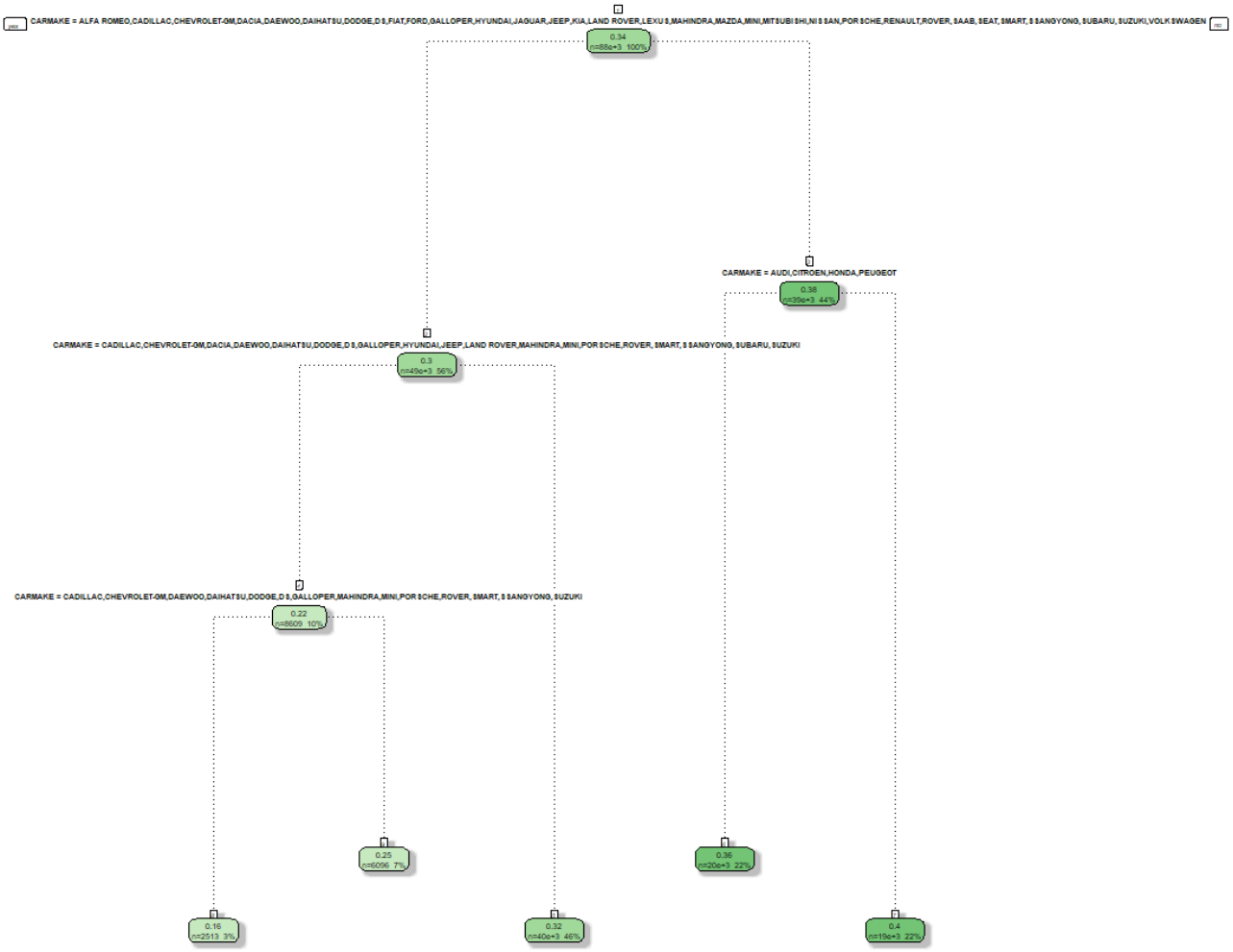


Figura 7.4: Árbol de clasificación para la marca del vehículo

Finalmente, se ha agrupado a partir del criterio de la distancia de *Gower* usando excepcionalmente el juicio experto y el conocimiento del mercado de modo análogo a como se hace en (Coskun, 2016). Finalmente se agrupan las marcas en torno a los siguientes tres centroides y estos son los resultados:

Grupo 1	Grupo 2	Grupo 3
ABARTH	BUICK	AUDI
ALFA ROMEO	DAEWOO	BMW
CHEVROLET-GM	DAIHATSU	CADILLAC
CITROEN	DS	CHRYSLER
DACIA	GALLOPER	DODGE
FIAT	MAHINDRA	JAGUAR
FORD	MINI	LANCIA
HONDA	NISSAN	MERCEDES
HUMMER	PGO	PORSCHE
HYUNDAI	ROVER	SUBARU
JEEP		TATA
KIA		TOYOTA
LAND ROVER		
LEXUS		
MAZDA		
MCC		
MITSUBISHI		
OPEL		
PEUGEOT		
RENAULT		
SAAB		
SEAT		
SKODA		
SMART		
SSANGYONG		
SUZUKI		
VOLKSWAGEN		
VOLVO		

Tabla 7.1: Clasificación de los vehículos según la frecuencia siniestral

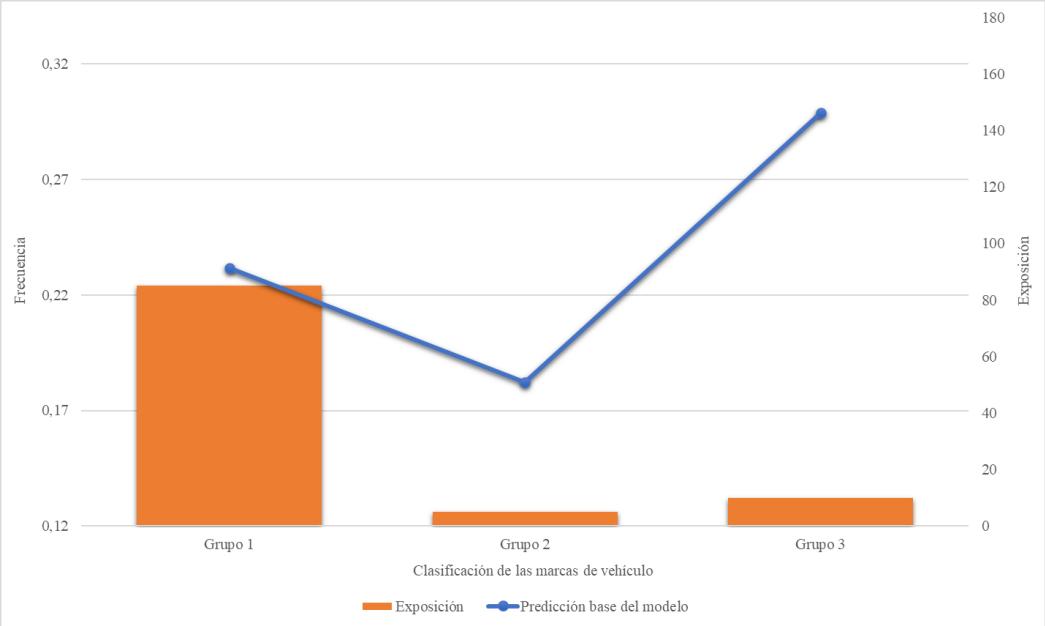


Figura 7.5: Tramificación de las marcas usando la distancia de *Gower*

3. Antigüedad del vehículo

La frecuencia derivada de la antigüedad del vehículo muestra una tendencia descendente a medida que la edad del mismo aumenta tal y como se observa en el siguiente gráfico.

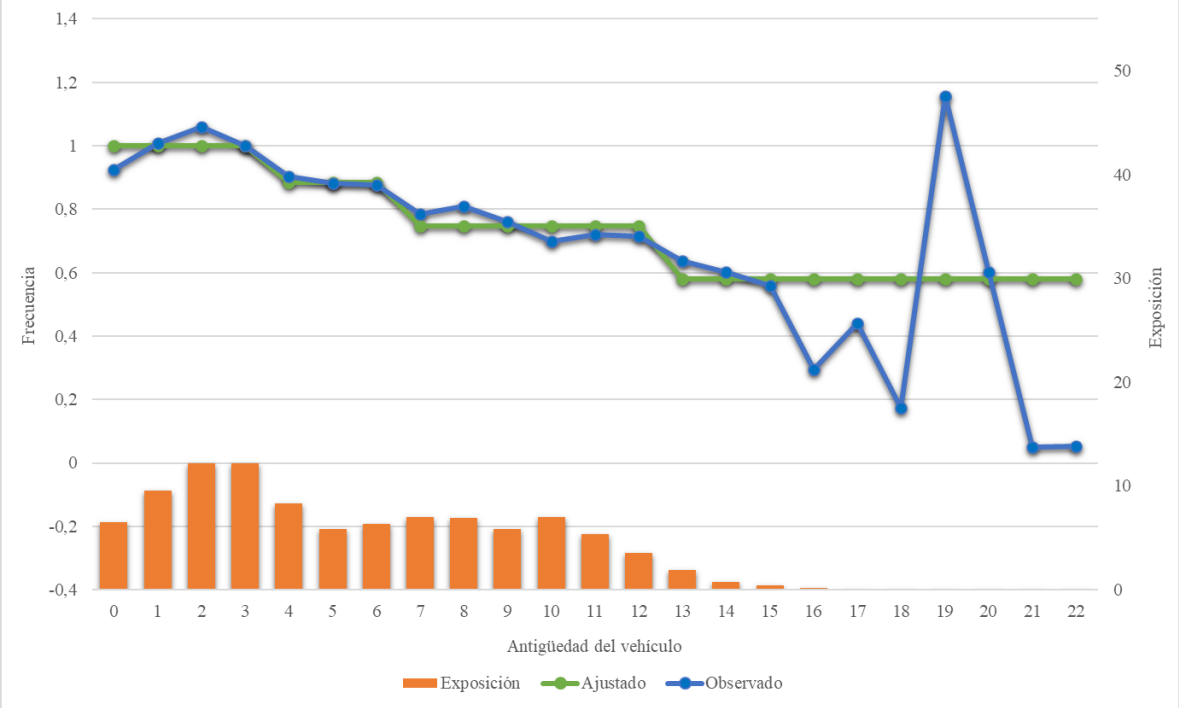


Figura 7.6: Distribución de la frecuencia según la antigüedad del vehículo

Por un lado, se aprecia que se tiene poca exposición para los niveles más altos, por lo que se ha agrupado ese tramo de antigüedades superiores a los 12 años, lo cual armoniza la tendencia y la explicabilidad de la variable en el modelo de frecuencia disminuyendo la volatilidad de las observaciones agrupadas.

Dadas las restricciones que impone el número de datos para poder hacer modelos con mayor capacidad predictiva se han realizado agrupaciones para niveles similares de frecuencia. El ejemplo del gráfico anterior ilustra el resultado del modelo observado contra el esperado agrupado en cuatro tramos.

Este tratamiento de agrupación se ha hecho para distintas variables en cada uno de los modelos GLM que se proponen. Los resultados del ajuste de los niveles se presentan en el siguiente gráfico.

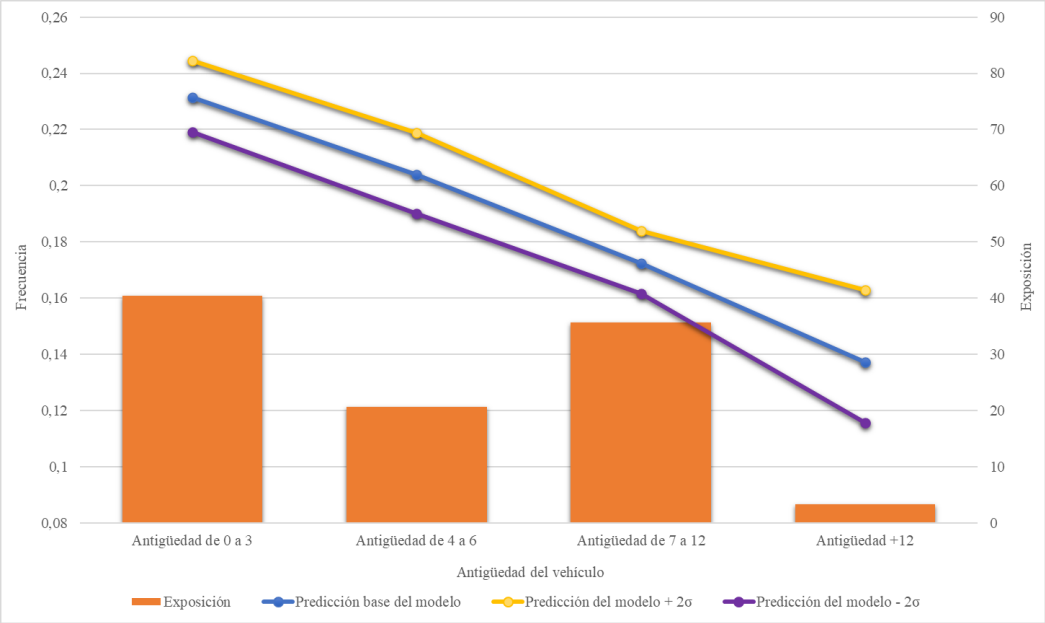


Figura 7.7: Predicción de la variable Antigüedad del vehículo en cuatro tramos

Además de los análisis univariantes que se ha realizado para las variables candidatas para considerar la inclusión de las variables, se realizan los gráficos de la estimación con la variable y los errores estándares[32]. En este caso se elige la variable dado que los errores estándar son estrechos y la estimación central de un nivel se sale de los intervalos de confianza de los niveles adyacentes.

[32] Los gráficos para todas las variables están en los anexos de este trabajo.

4. Número de puertas del vehículo

Otra de las variables disponibles en la base de datos es el número de puertas del vehículo. A partir de un criterio de razonabilidad, es de esperar que el número de puertas no tenga influencia en la frecuencia siniestral de los vehículos. Para confirmar esta hipótesis, se ha analizado esta variable con el fin de determinar si el número de puertas tiene un efecto en la frecuencia.

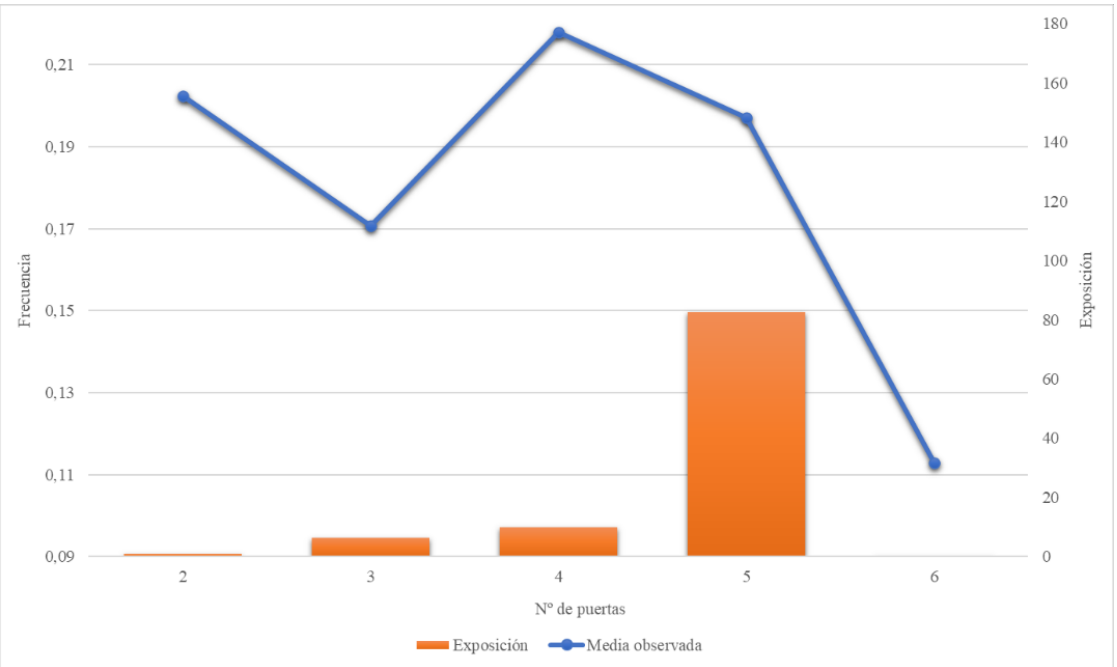


Figura 7.8: Distribución de la frecuencia según el número de puertas del vehículo

En el gráfico anterior no se aprecia una tendencia clara en cuanto a cómo afecta el número de puertas a la frecuencia siniestral de los vehículos. El comportamiento esperado carece de sentido aparente y por lo tanto esta variable no se ha incluido en el modelo.

5. Peso del vehículo

La siguiente variable analizada es un claro ejemplo de una variable numérica con una tendencia bien diferenciada, pero que presenta mucho ruido en las colas debido a la poca exposición. Por ello, estas variables numéricas también se han agrupado en tramos que añaden parsimonia y robustez al modelo simplificando el número de niveles que la variable toma. La variable presenta crecimiento con el valor del peso por lo que pese al excesivo número de niveles es clara candidata para entrar en el modelo.

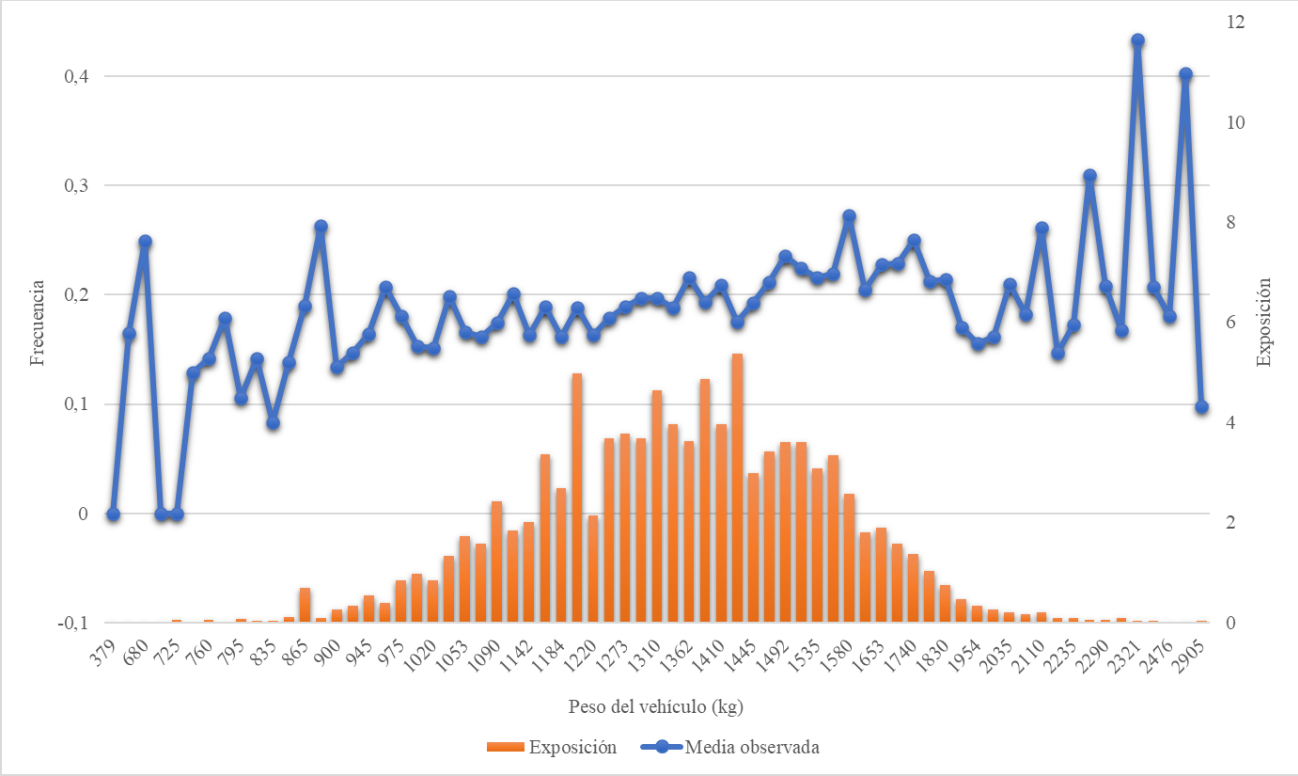


Figura 7.9: Distribución de la frecuencia según el peso del vehículo

6. Ayudas a la conducción (sistemas ADAS)

Con el objetivo de determinar si una variable debería incluirse o no en el modelo, se compara si la modelización incorpora ya el efecto de esa variable. En el caso de la variable CARGPS no está explicada por el modelo actual, tal y como se aprecia en el siguiente gráfico.

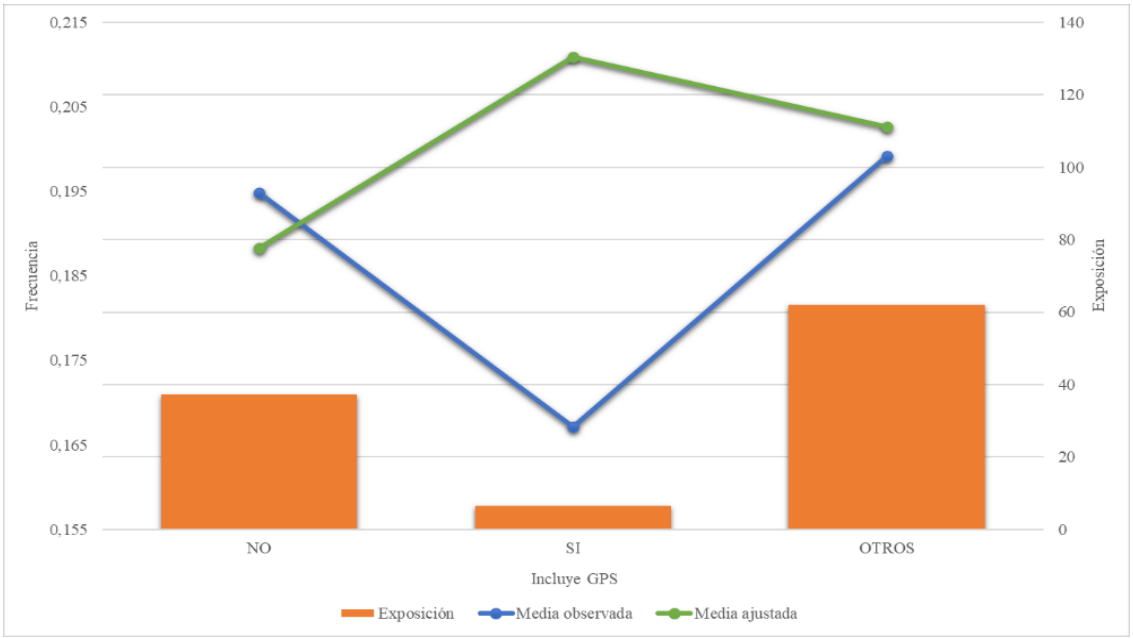


Figura 7.10: Efecto de la presencia del GPS en el vehículo

En el modelo que no incluye la variable se esperaría una mayor frecuencia para aquellos vehículos que si tienen GPS, mientras que en realidad tienen una frecuencia observada significativamente menor que aquellos vehículos que no tienen GPS o se desconoce si tienen.

Esta variable podría interpretarse como una aproximación para identificar que el vehículo está equipado con sistemas tecnológicos de ayuda a la conducción, también conocido como ADAS o Sistemas Avanzados de Asistencia a la

Conducción[33]. Estas mejoras tecnológicas están optimizando la seguridad en la conducción, lo que conlleva a una reducción de la siniestralidad y podría suponer una reducción del 57% de los siniestros en carretera como expresa un informe de la DGT (El Economista, 2018). En un futuro las entidades recogerán tanto el número de ayudas de a la conducción que incorporan los vehículos como la calidad de la ayuda para la prevención de accidentes.

Teniendo en cuenta ambas consideraciones se decide incorporar la variable que informa sobre la existencia de GPS al modelo de *scoring* GLM definitivo.

7.5.1.3 Construcción recursiva de los modelos

A la hora de modelizar, se debe tener en cuenta que existe una relación entre poder predictivo del modelo y su constitución lo más simple posible en cuanto a parámetros para mayor robustez, esto último se conoce como el Principio de parsimonia.

Tradicionalmente, se ha usado la Devianza como métrica principal para seleccionar el mejor GLM. Esta métrica se interpreta como los residuos del modelo, el problema es que su resultado mejora con la inclusión de más variables, por lo que usar únicamente esta métrica puede producir una complejidad perjudicial en el modelo final.

Siguiendo la metodología que propone Anderson et al. (2007) se ha utilizado una metodología *stepwise* en la selección de variables:

1. El criterio de entrada de las variables se realiza a partir de modelos GLM univariables para todos los posibles regresores y se elige el que minimiza la Devianza.

[33] La base de datos no incorpora información sobre otros sistemas de ayuda a la conducción.

2. Partiendo de ese modelo inicial se prueba con todas las demás variables, una a una usándose criterios de selección de modelos para determinar si son significativas o no. Las variables que se incorporan, además de presentar errores estándares bajos, deben presentar consistencia temporal para identificar si las tendencias son estables, y su inclusión mejora los modelos bajo los criterios de minimización del AIC y el BIC.

3. El proceso de selección del punto anterior se repite de forma recursiva hasta que la inclusión de nuevas variables no aporta más valor (minimiza el AIC y BIC).

Por lo tanto, para la selección de las variables que se consideran en el modelo final se han utilizado criterios de selección de modelos como el criterio de información de Akaike y el criterio de información Bayesiano. Para ambas técnicas, cuanto menor sea la puntuación de la métrica, mejor será el modelo. Se definen a continuación:

$$AIC = -2\ ln\ l(Y_i, \mu_i) + 2p$$

$$BIC = -2\ ln\ l(Y_i, \mu_i) + p\ log(n)$$

Donde “p” es el número de parámetros usados en el modelo, por su signo positivo empeora el resultado a medida que se aumentan las variables en el modelo, lo que permite detectar claramente si la variable aporta lo suficiente en cuanto a poder predictivo, para justificar su inclusión. El BIC a su vez tiene en cuenta el número de parámetros de los que dispone la muestra.

Una vez realizado el anterior proceso el mejor modelo en este análisis es el segundo y por tanto ha sido el finalmente seleccionado, dado que el último modelo con más variables no supera el modelo anterior según los criterios de información propuestos.

Modelos	AIC	BIC
CARFUEL + CARGPS + CARAGE + CARPVP	37.018,38	37.121,63
CARFUEL + CARGPS + CARAGE + CARPVP + CARMAKE	36.931,41	37.053,43
CARFUEL + CARGPS + CARAGE + CARPVP +CARMAKE + CARCC	37.577,71	37.764,86

Tabla 7.2: Criterio de selección de modelos mediante el AIC y el BIC

7.5.1.4 Creación del scoring con los parámetros de la regresión

Para el GLM[34], se ha propuesto un *scoring* que funciona estableciendo una puntuación usando los resultados de las betas de la regresión y su valor mínimo para cada variable. Este valor se corrige por un factor de significancia δ, que en este caso será 0,05. Esto es:

$$Scoring = \frac{\beta_{i,j} - MIN(\beta_{1,j} : \beta_{n,j})}{\delta}$$

Esta metodología arroja un resultado (scoring) para cada nivel de cada variable incluida en el modelo final, que incluye las variables de Edad del vehículo, Tipo de combustible, Precio del vehículo, Presencia de GPS y Marca.

7.5.2 Creación del scoring de vehículo a partir de GBMs

Como alternativa a los modelos de clasificación de vehículos con un GLM, se ha procedido a hacer el *scoring* de los vehículos usando un enfoque distinto, usando una técnica de *Machine Learning* de la familia *Boosting* como es el *Gradient Boosting Machine*. Usando la clasificación depurada anteriormente se ha procedido

[34] Esta metodología es la que se propone en el software propietario EMBLEM para la realización de *scoring*.

a clasificar los resultados en percentiles de 2,5% y se procederá a comparar los resultados usando métricas utilizadas para comparar la bondad de los modelos.

La ventaja de esta metodología, si se compara con otros algoritmos similares como el *XGBoost*, es que no requiere de un tratamiento previo de los datos. Asimismo, el tiempo de computación es más reducido que con algunas de estas otras técnicas.

Para la calibración del algoritmo, se separan los datos en entrenamiento y testeo, al igual que se hizo para el modelo GLM. Un 80% se dedicará a la calibración, mientras que el 20% restante se usará para analizar la precisión. Es decir, se han usado 70.435 registros para el entrenamiento y 17.609 para la fase de test.

Se han usado los paquetes de software libre R GBM, Caret y H20 para el uso de este algoritmo.

Con respecto al apartado de *tuning*, se ha realizado una búsqueda de la combinación de los hiperparámetros óptimos, algunos de los más importantes son:

- Número de árboles = Restringe el número de árboles totales, una medida fundamental con el fin de controlar el *overfitting*.

- Profundidad de los árboles = Se define como “d” y representa el número de particiones del árbol.

- Tasa de aprendizaje = Controla la velocidad por la que el algoritmo realiza el descenso del gradiente, un número bajo ralentiza significativamente el tiempo de ejecución, pero un número alto tiende a dar peores resultados no llegando a encontrar el mínimo global.

- Submuestreo = Permite que de forma aleatoria no se incluyan todos los datos del entrenamiento en cada una de las iteraciones, esto permite incorporar un componente estocástico a esta técnica.

Cabe recordar que todas las variables categóricas tienen que transformarse a factores para que el algoritmo converja correctamente. Además, se ha usado la técnica *K-Fold Cross Validation* para mejorar el modelo y se ha usado diez veces en el entrenamiento del modelo.

Asimismo, con fines de reproducibilidad, se ha fijado una semilla que permite replicar los resultados por terceras partes.

Los resultados se han dividido usando cuantiles homogéneos de 2,5%, para finalmente disponer de cuarenta subgrupos homogéneos.

7.5.3 Análisis comparados del *scoring* con metodologías GLM y GBM

Una vez realizada la clasificación de vehículos para la variable de frecuencia de daños propios con las dos metodologías se comparan mediante el uso de métricas de clasificación de modelos.

Estas métricas son muy similares y sirven para comparar las estimaciones de los modelos contra las observaciones reales. Estas medidas se utilizan para examinar los residuos de los modelos, ya que permiten identificar como de precisas son las predicciones realizadas y son muy utilizadas en regresión (Wesner, 2016).

Se han considerado inicialmente las siguientes métricas distintas. Estas son:

1. Error cuadrático medio (MSE)

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

2. Raíz del error cuadrático medio (RMSE)

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

3. Raíz del error cuadrático medio (MAE)

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

La segunda métrica presentada es una combinación lineal de la primera, por lo que se va a orientar el análisis a partir del RMSE y MAE únicamente.

El RMSE es la desviación cuadrática media del modelo. Mide las diferencias entre los valores del modelo o estimadores y los valores observados reales.

El MAE mide la magnitud media de los errores en un conjunto de predicciones, sin considerar su dirección. Es la media del valor absoluto de la diferencia entre la predicción y la observación del modelo. Al no incluir el término cuadrático todas las diferencias individuales tienen el mismo peso.

Ambas métricas toman valores en el rango [0,∞). Además, cabe destacar que se tratan de medidas con orientación negativa, lo que significa que cuánto menores sean los valores de las puntuaciones, mejores serán las proyecciones. La principal diferencia es que el RMSE perjudica los errores grandes y por ello suele ser la métrica más utilizada[35].

Con el objetivo de determinar que técnica de *scoring* se empleará en el siguiente modelo de frecuencia y severidad, como predictor, se comparan los dos modelos con las métricas anteriormente señaladas.

Los resultados del análisis se muestran en la tabla que aparece a continuación:

[35] Como por ejemplo en el trabajo de Guillen y Pesantez (2018), donde se utiliza este criterio del RMSE para comparar modelos de *Machine Learning* en la medición de la frecuencia en los seguros de automóviles.

Técnicas	RMSE	MAE
Scoring GLM	2,595	0,484
Scoring GBM	2,570	0,479

Tabla 7.3: Comparativa entre las dos técnicas de *scoring*

A la vista de los resultados obtenidos se observa como el modelo GBM clasifica ligeramente mejor la frecuencia de la siniestralidad de daños propios. Adicionalmente se ordena en 10 grupos de modo creciente la frecuencia media en función de cada uno de los métodos comparados.

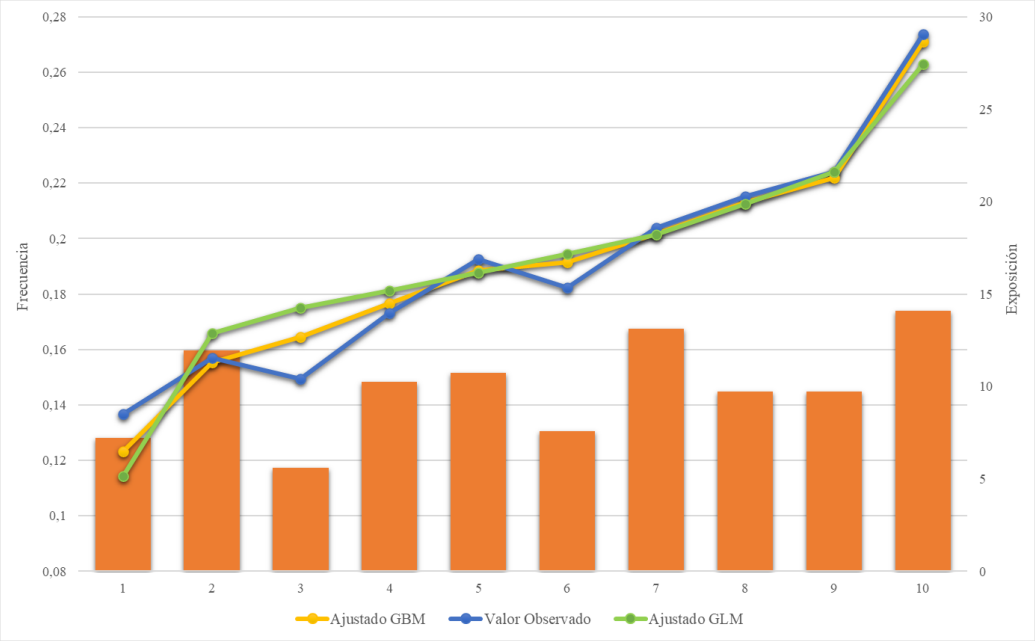


Figura 7.11: Resultados comparados del *scoring* GLM vs. GBM

Aunque los resultados no son rotundamente concluyentes, la agrupación según el modelo GBM clasifica mejor la frecuencia de daños propios. Teniendo en cuenta ambos resultados se escoge mejor técnica de clasificación el GBM construido a partir de las variables de vehículo. La combinación de esta técnica con los GLMs de frecuencia y severidad convierte a estos últimos en modelos híbridos.

7.6 Modelo de prima pura

La fórmula más extendida en la realización de modelos de tarifa en el seguro de automóviles es aquella que combina los modelos de prima pura, desglosando entre frecuencia y severidad que se combinan. En este caso se realizan por tanto dos modelos GLM distintos.

7.6.1 Selección de la función de enlace y la estructura de error

Para el modelo de frecuencia se ha considerado un modelo con una función de enlace logarítmica, con una estructura de error Poisson.

Para el modelo de severidad se ha considerado un modelo con una función de enlace logarítmica, con una estructura de error Gamma.

El uso de la función de enlace log en ambos modelos permite el uso de estructuras multiplicativas de tarifa que son fácilmente trasladables a los sistemas operacionales de emisión de pólizas de las entidades aseguradoras, por lo que son muy utilizados en la práctica aseguradora.

7.6.2 Análisis exploratorio inicial

En primer lugar, para cada modelo, se ha realizado un análisis de las correlaciones entre las variables. Si no se identifican y existe alta correlación pueden producirse casos de multicolinealidad, que generan predicciones muy inestables. Cuando se han detectado varias variables correlacionadas fuertemente, se ha analizado cual es la que aporta mayor poder explicativo y se han excluido las demás.

En los anexos adjuntos a este trabajo se muestra los gráficos univariables de las principales variables seleccionadas como candidatas a entrar en los modelos para la frecuencia y la severidad.

7.6.3 Construcción recursiva de los modelos

Como en el caso de la creación del *scoring* de vehículo para los modelos de prima pura se ha seguido la misma metodología s*tepwise*, propuesta en Werner (2005) para seleccionar las variables que se usarán en los modelos[36].

En cuanto a la validación de estos, se han analizado los residuos de ambos modelos, para ello se han usado los residuos de la devianza. Este chequeo es muy útil, ya que se pueden identificar problemas con las asunciones de los modelos o problemas de sobreajuste.

A continuación, se realizan los modelos individuales de frecuencia y severidad usando GLMs.

7.6.4 Modelo de frecuencia

En relación a la frecuencia siniestral se cuentan con 88.044 pólizas disponibles en el histórico de la compañía, únicamente el 10,32% o 9.087 pólizas han tenido al menos un siniestro. El objetivo es predecir la relación de la frecuencia siniestral con respecto a los factores de riesgo disponibles.

Además, se ha modelizado con un componente offset de la exposición de la variable.

En primer lugar, la fracción representa la frecuencia siniestral y f(X) es una combinación lineal de las variables independientes.

$$\frac{Numero\ de\ siniestros}{Exposicion} = e^{f(X)}$$

[36] Los regresores considerados con sus estimaciones y los errores estándares se pueden consultar en los anexos adjuntos.

Se toman logaritmos neperianos de ambos lados, tal que:

$$log\left(\frac{Numero\ de\ siniestros}{Exposicion}\right) = f(X)$$

Se separa la fracción usando las propiedades de los logaritmos, para dejar la variable respuesta, el número de siniestros, a un lado de la ecuación.

$$log\,(Numero\ de\ siniestros) - log\,(Exposicion) = f(X)$$

Finalmente, se identifica el componente offset agregado a f(X).

$$log\,(Numero\ de\ siniestros) = f(X) + log\,(Exposicion)$$

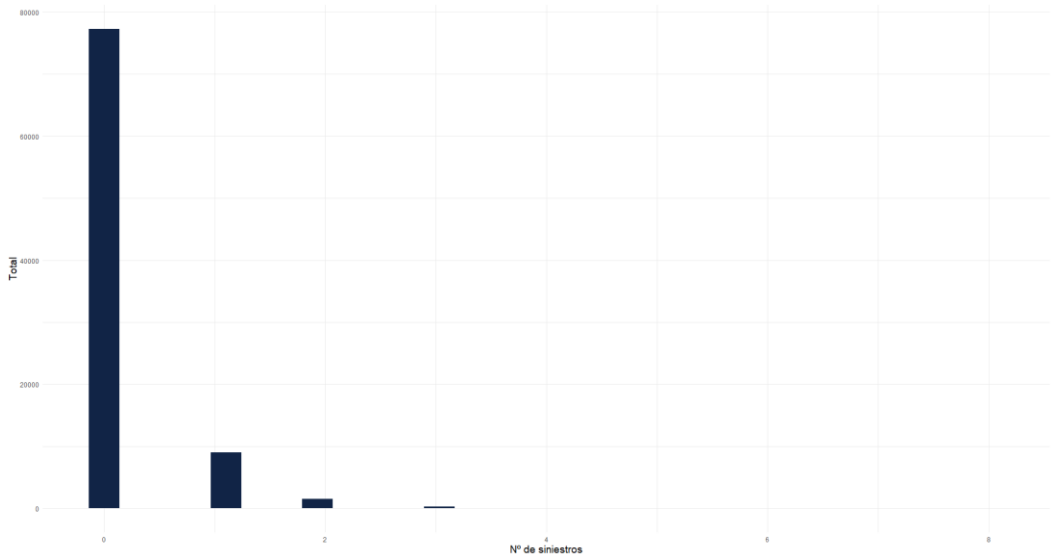


Figura 7.12: Distribución del número de siniestros de la compañía

En el gráfico anterior se aprecia una asimetría fuerte positiva, donde el mayor peso de esta distribución se encuentra en cero. Luego se observa una cola a la derecha que se extiende hasta el valor ocho. La forma de los datos observados se corresponde con una distribución de Poisson.

El modelo final seleccionado incluye las variables edad del tomador, tipo de producto, el *scoring* GBM de vehículo, el Bonus/Malus y los años de antigüedad en la compañía.

Con el objetivo de cerciorarse de que las asunciones del modelo son correctas y no existen desequilibrios en el modelo, se ha realizado un análisis de los residuos estandarizados de la devianza.

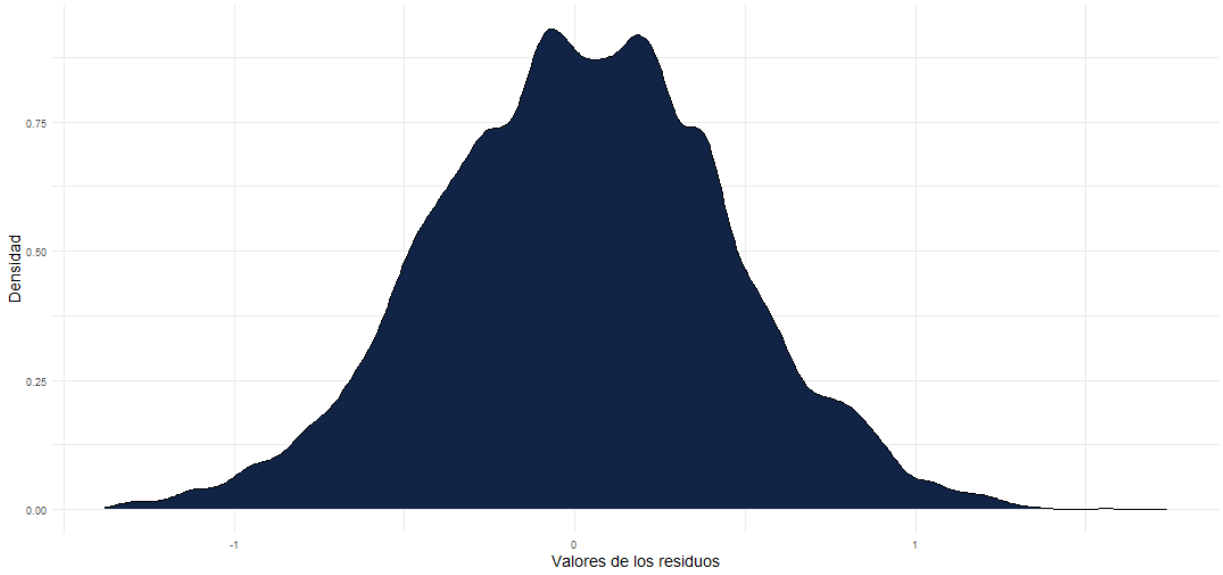


Figura 7.13: Distribución de los residuos para el modelo de la frecuencia

La densidad de los residuos refleja que la variabilidad a lo largo del modelo es relativamente constante y homogénea, con su media en torno a cero. Esto sugiere que la función de la varianza seleccionada es apropiada y el modelo es válido.

7.6.5 Modelo de severidad

En cuanto al modelo de coste, se ha seguido la metodología propuesta en (Tiwari,2020) para el procedimiento de modelización de la severidad en seguros.

Para el tratamiento estadístico se ha filtrado por aquellas pólizas que han experimentado al menos un siniestro a lo largo del año. Se observa cierta homogeneidad, con muy pocos valores atípicos y de importe relativamente bajo. El 99,55% de los siniestros no superan el umbral de 10,000 euros. No ha sido necesario eliminar ningún siniestro.

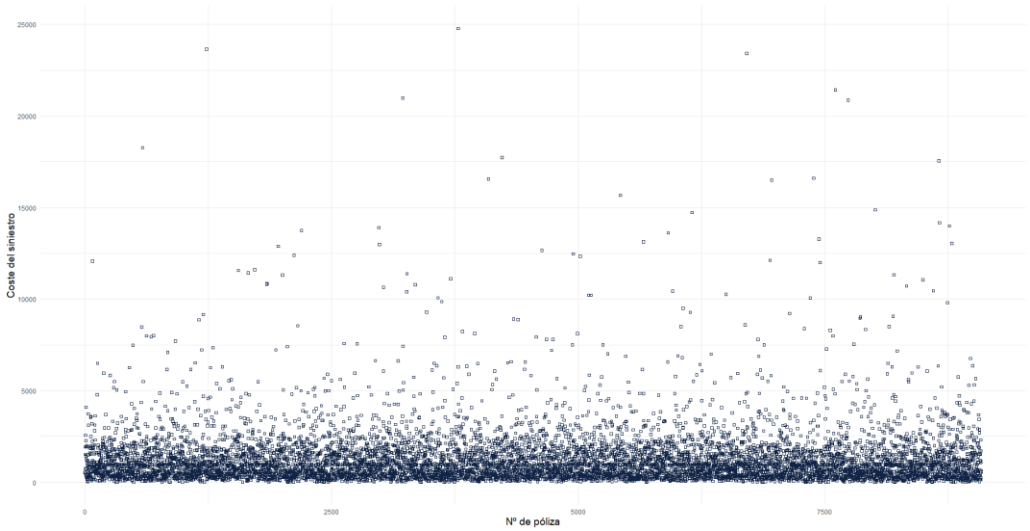


Figura 7.14: Cuantía de los siniestros por póliza

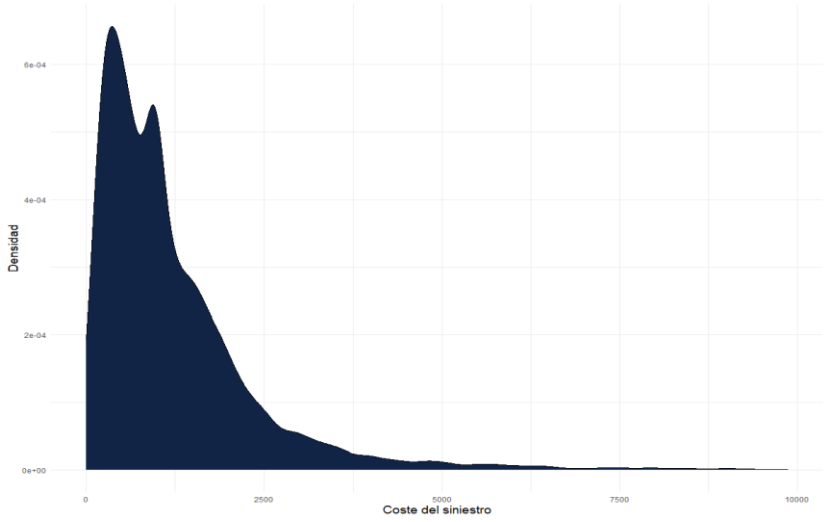


Figura 7.15: Distribución del coste medio

Se aprecia una distribución con asimetría positiva, teniendo la mayor parte de siniestros una cuantía entre 0 y 1.000 euros y una cola larga a la derecha de la distribución. Esta indicación de asimetría positiva de una variable continua sugiere que la modelización a partir de una Gamma es una propuesta acertada.

Para este modelo, la variable objetivo es la cuantía de los siniestros. La función enlace que relaciona las variables independientes con la variable objetivo es la logarítmica como ya se señaló anteriormente. Esta transformación permite obtener la estructura multiplicativa que facilita la implementación operativa, pese a que la teoría propone el uso de la función recíproca.

Además, se ha modelado con el término *offset* de la transformación logarítmica del número de siniestros. A continuación, se explica su derivación.

Inicialmente la estructura del GLM es la siguiente, dónde la fracción representa el coste medio por siniestro y f(X) es una combinación lineal de las variables independientes.

$$\frac{Coste\ del\ siniestro}{Numero\ de\ siniestros} = e^{f(X)}$$

Se toman logaritmos neperianos de ambos lados, tal que:

$$log\left(\frac{Coste\ del\ siniestro}{Numero\ de\ siniestros}\right) = f(X)$$

Se separa la fracción usando las propiedades de los logaritmos, para dejar la variable respuesta, coste del siniestro, solo a un lado de la ecuación.

$$log\,(Coste\ del\ siniestro) - log\,(Numero\ de\ siniestros) = f(X)$$

Finalmente, se identifica el componente offset agregado a f(X).

$$log\,(Coste\ del\ siniestro) = f(X) + log\,(Numero\ de\ siniestros)$$

El modelo final seleccionado incluye las variables edad del tomador, el tipo de producto y el *scoring* de vehículo. Finalmente, una vez calibrado el GLM, se muestra una distribución de los residuos estandarizados para el modelo de severidad.

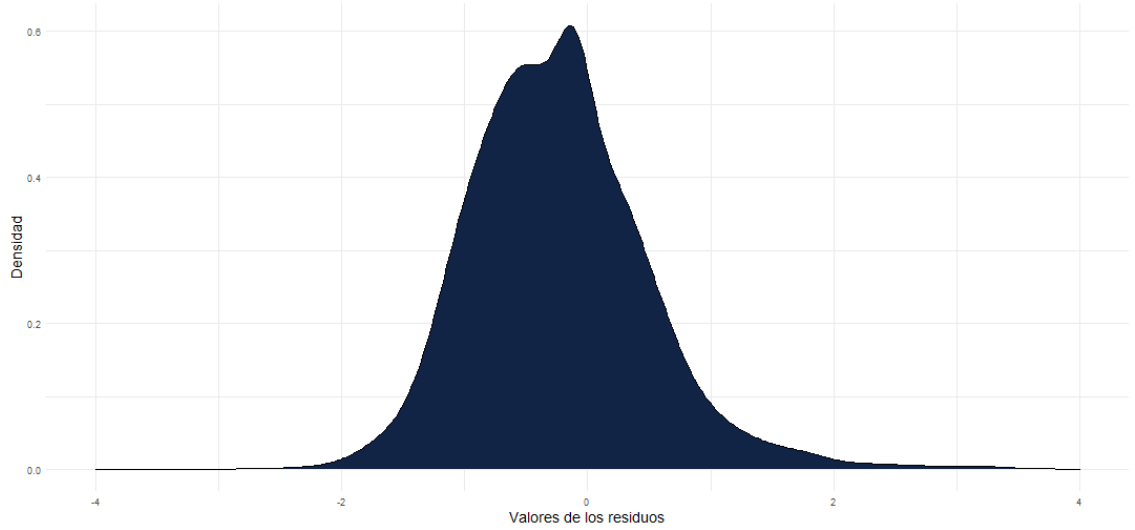


Figura 7.16: Distribución de los residuos para el modelo de coste medio

La densidad de los residuos refleja que la variabilidad a lo largo del modelo es relativamente simétrica, pese a presentar una ligera desviación hacia la izquierda. Esto apunta a que la función de la varianza seleccionada es apropiada y el modelo es válido.

7.6.6 Resultados modelo prima pura

Las relatividades de los modelos de frecuencia y severidad, derivadas a partir de las betas transformadas, se encuentran recogidas en la siguiente tabla. Todas las variables son significativas individualmente y sus coeficientes tienen un impacto relevante sobre las respectivas variables respuestas.

Variables	Niveles	Relatividades frecuencia	Relatividades severidad
Nivel base	-	0,217	1128,950
Edad del asegurado	<27	1,211	1,454
Edad del asegurado	[27 - 52]	1,000	1,000
Edad del asegurado	[53 - 70]	1,144	0,927
Edad del asegurado	>70	1,538	1,007
Scoring	<19	0,946	0,958
Scoring	[19 - 22]	1	1
Scoring	23	1,05	1,013
Scoring	24	1,146	1,028
Scoring	25	1,161	1,029
Scoring	26	1,226	1,032
Scoring	[27 - 28]	1,180	1,034
Scoring	29	1,369	1,038
Scoring	>29	1,374	1,041
Año de renovación	1	1,115	-
Año de renovación	2	1,148	-
Año de renovación	3	1,627	-
Año de renovación	4	1,086	-
Año de renovación	5	1,067	-
Bonus en renovación	<4	1,941	-
Bonus en renovación	[4 - 6]	1	-
Bonus en renovación	>=7	0,557	-
Tipo de producto	Daños propios con franquicia	1	1
Tipo de producto	Daños propios	3,145	1,024

Tabla 7.4: Relatividades resultantes de los modelos de frecuencia y severidad

La interpretación de estos resultados se entiende como que los niveles base son 0,217 y 1.128,95 euros para la frecuencia y la severidad respectivamente. Para la frecuencia hay un nivel con una relatividad asociada que recarga considerablemente el nivel base. Se trata del nivel del producto que cubre Daños Propios sin franquicia. Esto es razonable, ya que la franquicia tiene un efecto moderador sustancial del número de siniestros que se reportan, dado que proporciona un incentivo económico al tomador para prevenir la ocurrencia de los siniestros (Feldman y Brown, 2005). Adicionalmente, la franquicia conlleva un ahorro considerable de recursos destinados a atender los siniestros, otros gastos

administrativos y, por supuesto, la cuantía de las indemnizaciones (The Actuary, 2019).

El resto de las relatividades se aplican multiplicativamente como recargos o descuentos a imputar sobre el nivel base, tal que:

$$Frecuencia \times Coste\ medio = Prima\ pura$$

Finalmente, la distribución de la prima pura de la cartera es la siguiente:

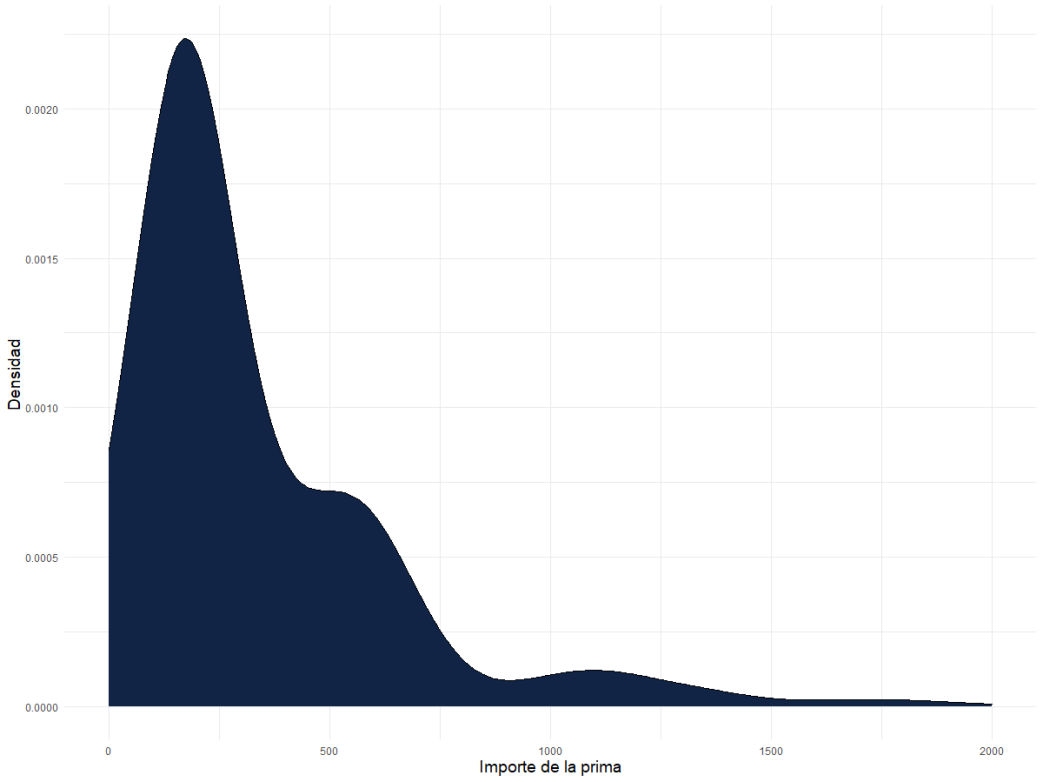


Figura 7.17: Distribución de las primas para la cartera para la Cobertura de Daños Propios

En este gráfico se puede apreciar como existen la forma de la distribución de las primas sugiere dos distribuciones distintas. Como ya se ha trasladado previamente existen diferencias notables entre el producto Todo riesgo con y sin franquicia.

Si se separan las primas por tipo de producto se obtienen los siguientes resultados:

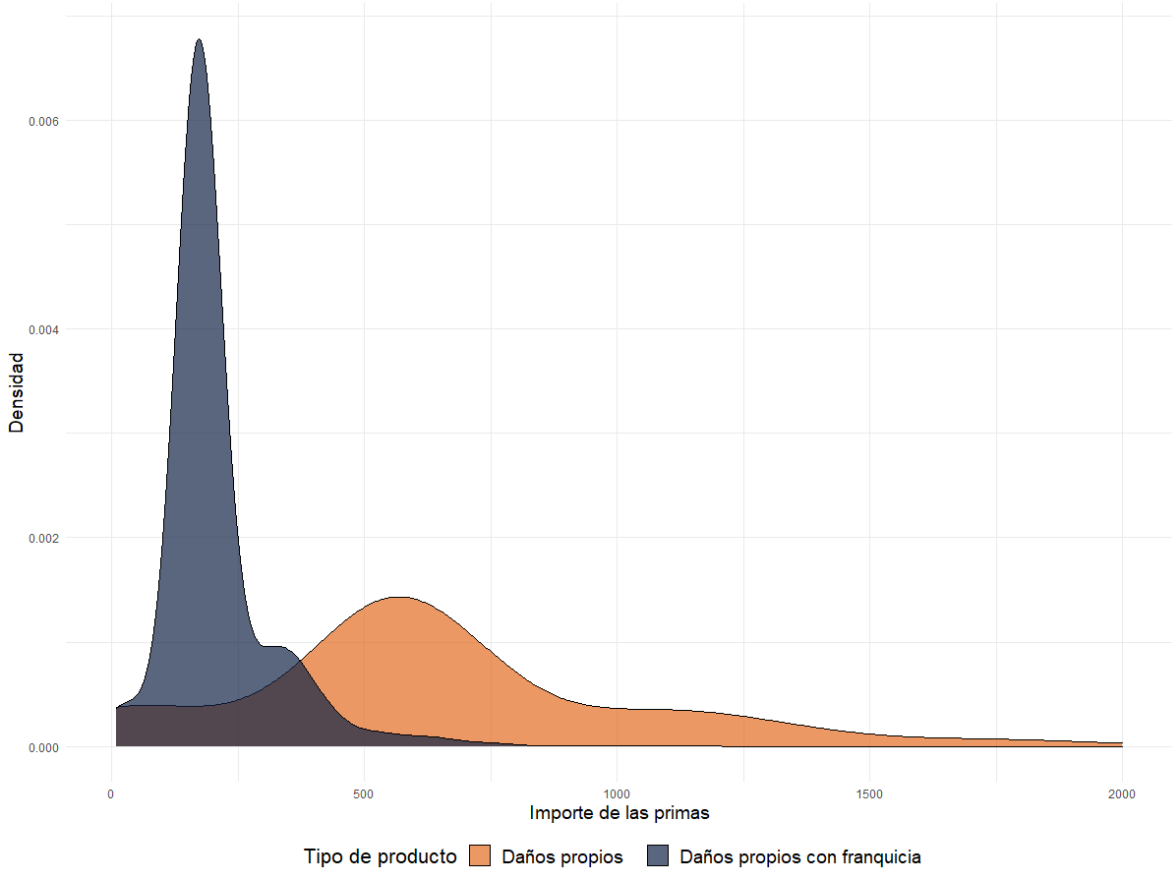


Figura 7.18: Distribución de las primas por tipo de producto

De los resultados del gráfico se concluye que, además de la diferencia sustancial en precio de los productos con franquicia y sin franquicia existe una variabilidad significativa derivada en las primas por las características de los riesgos asegurados. La variabilidad es mayor en el caso del producto sin franquicia.

7.7 Modelo de retención

Se ha realizado un modelo de retención para estimar la probabilidad de renovación de la póliza, teniendo en cuenta las características individuales del tomador y su comportamiento frente a la prima ofrecida. El modelo predice la probabilidad de anulación de la póliza, pero se obtiene la de retención mediante su complementaria tal que:

$$Retencion = (1 - Anulacion)$$

Con respecto a los datos utilizados, se ha empleado una submuestra de 20.713 pólizas. Estas variables se corresponden con pólizas en vigor en 2018 y se ha añadido la variable de cancelación en la renovación del año 2019, que se registró posteriormente en el 2019 determinando si se renovó o no la póliza.

A continuación, se destacan las principales variables:

- INDICADOR CANCELACIÓN: Es la variable respuesta, es una variable dicotómica 0/1 que indica si la póliza renovó o no.

- EDAD PRINCIPAL: Recoge la edad del tomador principal.

- PRIMA ANTERIOR: Registra la prima correspondiente del año anterior, aquella correspondiente a 2018.

- VARIACIÓN RELATIVA: Consiste en el cociente entre el precio propuesto y el actual. Esta es probablemente la variable más relevante, que permite cuantificar la elasticidad de los clientes.

- PROMEDIO COMPETENCIA: Se ha calculado la distancia absoluta entre el precio propuesto y la media de los tres primeros competidores.

- AÑOS DE ANTIGÜEDAD: Determina la antigüedad de la póliza en la compañía.

- TIPO DE PÓLIZA: De misma forma que para los modelos de frecuencia y severidad, esta variable indica la cobertura de la póliza.

7.7.1 Selección de la función de enlace y la estructura de error

Para el modelo de anulación de pólizas se ha considerado un modelo GLM con una función de enlace *logit* y una estructura de error Binomial. La derivación del modelo teórico resultando en la estimación de la probabilidad se encuentra en el apartado de modelos de este trabajo.

7.7.2 Análisis exploratorio inicial

Se ha realizado el análisis univariante de las variables siguiendo la metodología descrita en los modelos anteriores. Los resultados más relevantes del comportamiento de la anulación de los clientes se describen a continuación.

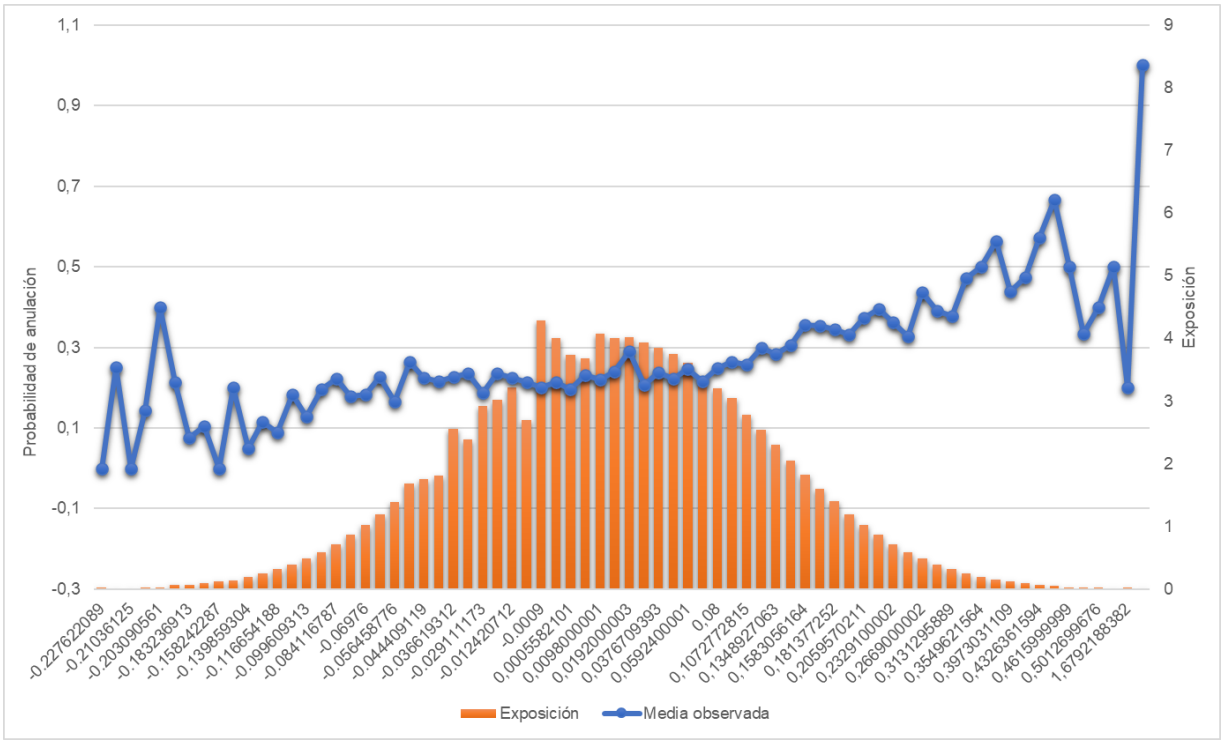


Figura 7.19: Distribución de la probabilidad de anulación según el cociente entre prima propuesta y prima en vigor

Este primer gráfico muestra la diferencia relativa entre la prima en vigor y la prima ofrecida en la renovación, midiendo por tanto la elasticidad precio demanda.

Se aprecia una tendencia lineal positiva con varios comportamientos diferentes. Hay una ligera pendiente positiva en la parte izquierda de la distribución. Esto sugiere que los clientes apenas reconocen ligeros descuentos de hasta un -5% aplicados sobre la prima, ya que no se aprecian diferencias significativas entre aquellos que reciben pequeños descuentos y los que reciben el mismo precio. El intervalo central, con hasta un 3% de variación donde no parece que la variación influya en la anulación. Para incrementos de prima superiores al 3% se incrementa la pendiente lo que significa que el cliente si considera la variación del precio a la hora de renovar su póliza.

Adicionalmente, se aprecia ruido en las colas que se agrupará para mayor robustez en la predicción.

La siguiente variable tiene también un efecto material en la modelización de la tasa de anulación.

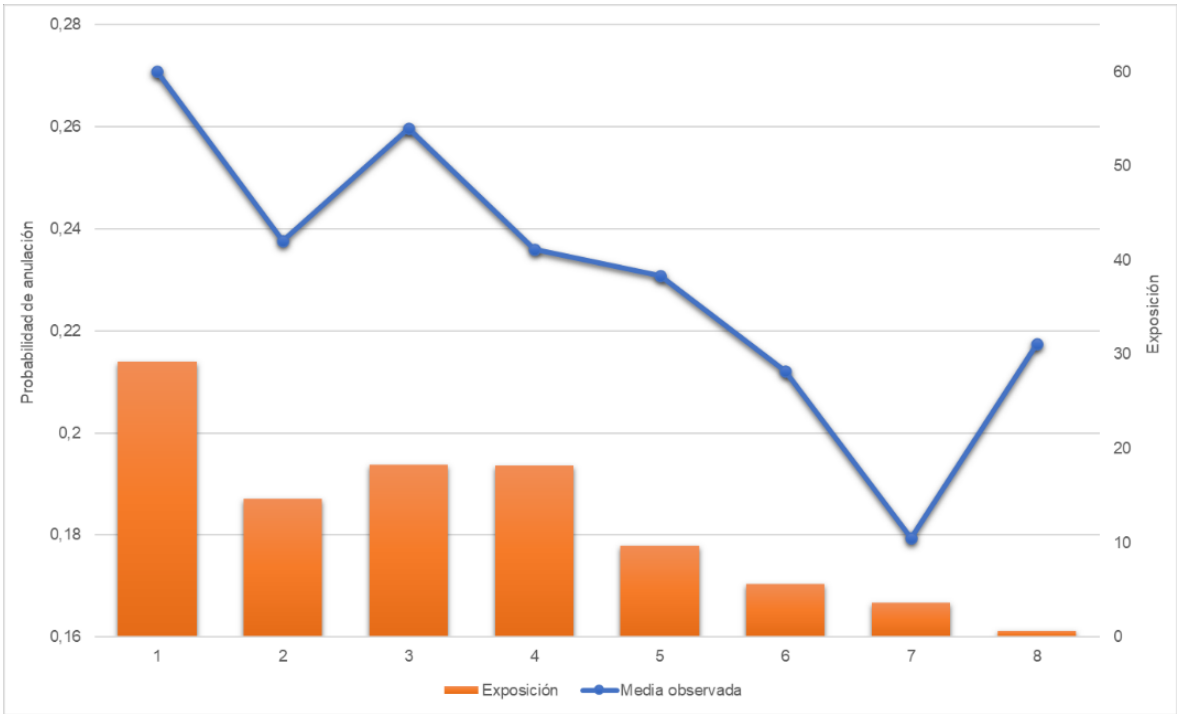


Figura 7.20: Distribución de la probabilidad de anulación según la antigüedad de la póliza

Un resultado significativo es que a medida que transcurren los años la póliza tiene menor probabilidad de anulación. Esto sugiere que el cliente es más reacio a cambiar de compañía cuanta mayor antigüedad tiene. Por lo tanto, el estudio de esta variable permite detectar los perfiles que son más inelásticos ante cambios de precio según su fidelidad a la compañía.

Con respecto a este punto, también es muy relevante el siguiente gráfico que se muestra a continuación.

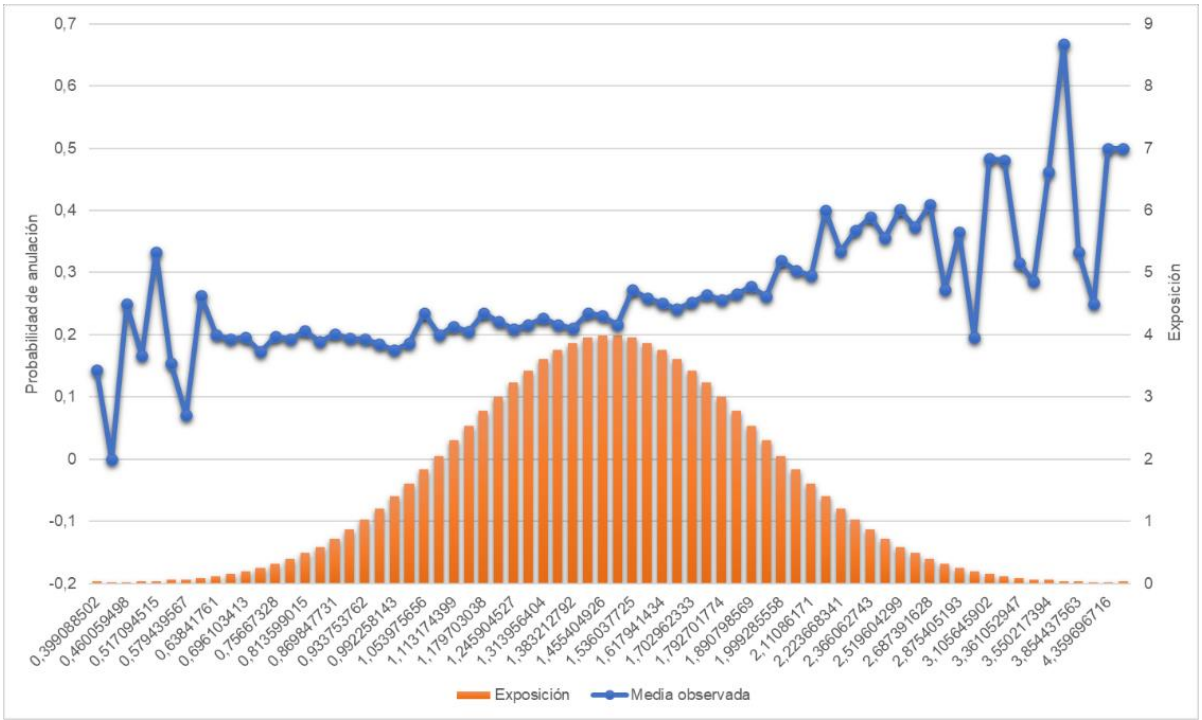


Figura 7.21: Distribución de la probabilidad de anulación según el cociente entre prima propuesta y la prima del promedio de la competencia.

El gráfico anterior representa como varía la probabilidad de anulación según la diferencia relativa con respecto al promedio de los tres primeros competidores. En la figura se observa una relación no lineal entre el posicionamiento de precios más elevados en la empresa, con una pendiente con tendencia exponencial a medida que la compañía aumenta las tarifas. Sin embargo, no se observa el mismo efecto en el caso contrario.

Como conclusión, la idea general sugiere que se pierden más clientes de la cartera al aumentar el precio de las primas por encima de los valores de los competidores. No obstante, si se observan detenidamente los resultados estos sugieren que una bajada de los precios no necesariamente implica una correspondiente disminución de la probabilidad de anulación.

Por tanto, se recomienda comprobar y monitorizar de forma regular el posicionamiento frente a los competidores.

7.7.3 Resultado del modelo

Para obtener la probabilidad final del modelo se deben agregar los parámetros de los regresores obtenidos a partir de la transformación de los predictores lineales cuya demostración se desarrolló en el capítulo de modelos estadísticos de este trabajo.

Asimismo, cabe recordar que este modelo es de cancelación, por lo que se ha obtenido la probabilidad de retención, que es la que se necesita a partir de su probabilidad complementaria.

Adicionalmente, se obtienen los datos de los competidores, usando la herramienta de *web-scraping* y, sobre los datos obtenidos, se ha derivado la tarifa de la competencia a partir de un modelo GBM[37]. Este resultado se utiliza como regresor en el modelo de cancelación comparando la prima media de los competidores con la prima ofrecida.

Para finalizar, las relatividades del modelo de cancelación se muestran a continuación.

[37] El código con la descripción del modelo GBM se incorpora en los anexos de este trabajo.

Variables	Niveles	Relatividades anulación
Nivel base	-	-1,3664
Edad del asegurado	< 28	0,2265
Edad del asegurado	>= 29	0,0000
Tipo de producto	Daños propios con franquicia	0,0000
Tipo de producto	Daños propios	0,2045
Diferencia relativa vs prima anterior	< -0.1	-0,5971
Diferencia relativa vs prima anterior	[-0.1, -0.05]	-0,1464
Diferencia relativa vs prima anterior	(-0.05, 0.02]	0,0490
Diferencia relativa vs prima anterior	(-0.02, 0.02]	0,0000
Diferencia relativa vs prima anterior	(0.02, 0.05]	0,0549
Diferencia relativa vs prima anterior	(0.05, 0.12]	0,1496
Diferencia relativa vs prima anterior	> 0.12	0,4905
Diferencia relativa vs promedio competidores	< 0.05	0,0000
Diferencia relativa vs promedio competidores	[0.05, 0.5]	0,1469
Diferencia relativa vs promedio competidores	[0.5, 1.1]	0,1469
Diferencia relativa vs promedio competidores	> 1.1	0,4280
Prima en vigor	< 230	-0,3739
Prima en vigor	[230, 460]	0,0000
Prima en vigor	> 460	0,3508

Tabla 7.5: Relatividades resultantes del modelo de anulación

Estas relatividades[38] se usarán en el siguiente apartado con el objetivo de derivar el precio óptimo para cada cliente. Cabe mencionar que la función de demanda correspondiente es monótona decreciente, requisito necesario para obtener una única solución para el problema de búsqueda de un máximo global para cada caso.

7.8 Algoritmo de optimización individual

La retención y la conversión de clientes se ha convertido en un campo de creciente atención por parte de los actuarios de tarificación durante los últimos 20 años. La generalización de los agregadores de precio, en los que es posible comparar primas no ha hecho sino exacerbar ese interés de los aseguradores. Es una práctica actuarial extendida internacionalmente modelar no sólo el coste del riesgo sino también la demanda de seguro. Para ello es necesario contar con variables

[38] En el caso de las variaciones relativas de prima, se ha procedido a suavizar e interpolar las variaciones de precio, en intervalos de 50 puntos básicos. Este ajuste permite tener un mayor número de estrategias de precio.

adicionales a las utilizadas en la tarificación clásica, como se ha visto en el apartado anterior.

El proceso de optimización de primas consiste en el uso de información sobre el comportamiento del cliente con el objetivo de determinar que prima se ofrece a cada individuo.

Una definición más formal es la que se realiza en CAS (2014), que expresa la optimización como "el complemento de los modelos tradicionales de coste para incluir modelos cuantitativos de demanda de clientes para fijar los precios de la tarifa. El resultado final es un conjunto de ajustes sobre los modelos de coste del riesgo por segmento de cliente para las diferentes clases de riesgo".

Por tanto, en la optimización se incluyen consideraciones sobre el comportamiento del cliente y del mercado, yendo más allá de una concepción clásica de tarificación basada en riesgo[39].

En la práctica la optimización requiere un exhaustivo proceso de modelado actuarial ya que requiere modelos de prima, de competencia, de conversión o retención y un algoritmo de optimización matemática que integre los anteriores modelos.

De las opciones posibles de modelado, ya referidas en el capítulo 5 de taxonomía de la tarificación, se opta por la optimización matemática individual a partir de la técnica *Grid Search*[40] (Earnix, 2020). Recibe este nombre porque permite ofrecer precios particularizados para cada asegurado según su perfil de riesgo y la propensión individual a renovar la póliza según las condiciones ofrecidas. Para cada cliente se optimiza la prima ofrecida en un rango discreto entre dos límites preestablecidos. En cada caso se selecciona la prima que maximice la función

[39] Existen no obstante algunas limitaciones en algunos países como algunos estados, ODI (2015), de Estados Unidos, donde se consideran estás técnicas como injustas. Para más detalle véase CAST (2015) y Hartwig (2015).

[40] El problema de optimización matemática podría resolverse a partir de una aproximación analítica. Sin embargo, aplicar esa aproximación supone aplicar asunciones que se quieren relajar como para los casos en los que la función a optimizar es discontinua o que no sea convexa.

objetivo, en este caso el margen esperado[41], analizando los resultados obtenidos en todos los valores del rango.

Expresado matemáticamente, si se analiza el porfolio en su conjunto, el problema de optimización se define tal que:

$$max \sum_{i=1}^{n} VAN_i(p_i^*)$$

Con las consecuentes restricciones:

$$f_i(p) = C_i$$

$$I_i \leq p_i^* \leq S_i$$

Dónde:

- p_i^*: Precio optimizado para el cliente *i*.
- VAN_i: Valor actual neto del beneficio por póliza. Tradicionalmente es frecuente derivar esta expresión únicamente para un año, como en este trabajo, por lo que se considera un único flujo neto.
- $f_i(p)$: Restricciones aplicadas al modelo, en relación a la prima ofrecida para el cliente *i*.
- $I_i y\, S_i$: Límites superior e inferior que restringen la optimización, de modo cuantitativo para cada cliente *i*.

En el caso de que adicionalmente se incorporen restricciones globales este problema se resuelve utilizando los multiplicadores de Lagrange, una técnica

[41] Existe la opción de optimizar retención o una combinación de retención y margen. Son posibles múltiples combinaciones de enfoques dependiendo de la estrategia de la compañía, condicionado frecuentemente por el ciclo de mercado y los objetivos empresariales marcados.

algebraica que permite resolver problemas de búsqueda de un máximo o mínimo dadas una serie de restricciones.

En este proyecto no se fijan restricciones globales en el proceso de optimización, dejándose para futuras líneas de trabajo desarrollar adicionalmente estas restricciones.

Las restricciones individuales fijadas en este trabajo son las siguientes:

- Las variaciones máximas de la prima propuesta no podrán exceder +/- 10% por cliente respecto a la prima del año anterior.
- Dentro de los rangos establecidos se fijan variaciones de 50 puntos básicos.
- La prima mínima nunca podrá ser inferior al 75% de la prima comercial técnica derivada a partir de los modelos de riesgo[42] obtenidos en este mismo capítulo.

Desde un punto de vista matemático, el problema se plantea con el objetivo de encontrar aquella prima que maximiza la siguiente expresión para cada cliente:

$$M^* = max\left[P(\theta) \times (\theta - \kappa - \delta)\right]$$

Dónde:

- *M**: Margen optimizado para cada cliente.
- θ: Prima comercial ofrecida.
- P(θ): Probabilidad de retención dada una determinada prima ofrecida.
- κ: Costes de adquisición y administración de la póliza.
- δ: Prima pura.

[42] Dado que el modelo de riesgo se ha derivado sólo para la cobertura de Daños Propios se elevan los resultados obtenidos considerando el peso adicional que tienen el resto de coberturas según los datos obtenidos de (ICEA, 2021). La prima comercial se obtiene incorporando los recargos de gestión sectoriales obtenidos para el año 2019 y que se han extraído de (MAPFRE Economics, 2020).

7.8.1 Resultados del modelo de optimización

Operativamente los resultados de la optimización se han aplicado para una muestra de 20.712 pólizas. En este caso se han calculado las probabilidades de retención para cada póliza, teniendo en cuenta las características de cada cliente y considerando las variaciones de prima. Esto permite obtener la frontera eficiente de 41 puntos para cada una de las variaciones de prima, retención y margen por cliente[43], lo que supone finalmente calcular un total de 849.192 valores diferentes para cada una de las variables consideradas.

Además, con el objetivo de validar el uso de esta técnica, se han calculado las diferencias entre un escenario de primas optimizadas *versus* un escenario alternativo tradicional donde todas las pólizas reciben un aumento fijo del 1% con la consideración de una hipótesis fija de inflación del coste medio del producto[44].

Los resultados finales del modelo de optimización se pueden medir bajo un doble prisma. Por un lado, la comparación de las estrategias aplicadas para cada cliente de modo individual. Por el otro, a través del valor agregado para toda la cartera de clientes.

El caso concreto para tres clientes elegidos al azar, del total de la cartera analizada, se muestra a continuación. En primer lugar, se presenta el margen técnico por comparación entre la prima ofrecida y el coste del riesgo en euros para cada cliente.

[43] La visualización de las fronteras eficientes por cliente se ha realizado en la herramienta desarrollada en R Shiny.

[44] La aplicación de una asunción de este tipo es bastante frecuente en los procesos de renovación de tarifas de muchas compañías aseguradoras.

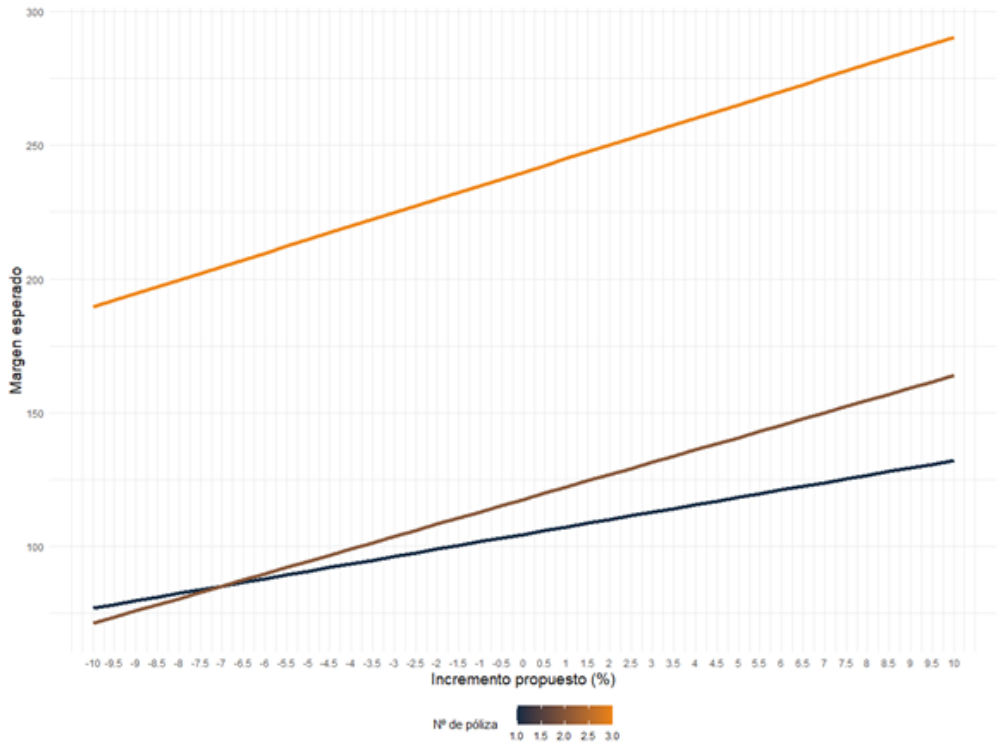


Figura 7.22: Margen técnico por cliente por variación relativa de prima

A la vista de los resultados obtenidos se observa como estos 3 clientes tienen márgenes positivos para todo el rango de variación consecuencia de que la prima ofrecida de la que se parte está por encima de la prima comercial técnica calculada incluso para los casos en los que se ofrecen hasta variaciones del -10% del precio del año precedente. Cada póliza presenta una aportación de valor diferente siendo el tercero de ellos el que aporta un valor significativamente mayor.

Como se puede apreciar, el margen técnico individual se caracteriza por ser una función lineal con pendiente positiva con respecto a las variaciones de precio ofrecidas, aunque con pendiente diferente en función de las características de cada póliza.

A continuación, se presenta las curvas con las probabilidades de retención por cliente en función de la variación de prima ofrecida al cliente.

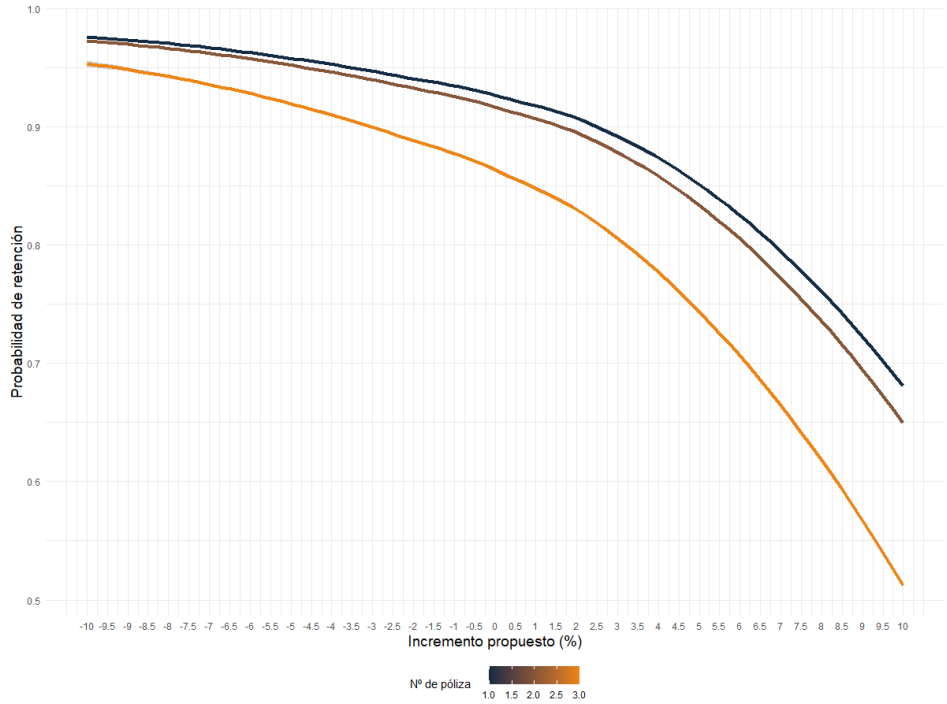


Figura 7.23: Probabilidad de retención por cliente y variación relativa de prima

A la vista de los resultados obtenidos se observa un comportamiento parecido en las curvas de retención para los dos primeros y clientes y distinto para el tercero. En cualquier caso, las distribuciones de probabilidad de retención, en función de la variación de la prima ofrecida, se caracterizan por tener una forma cóncava con valores de retención que comienzan a decrecer significativamente a partir de variaciones superiores al 4% en precio.

Finalmente se obtiene el valor optimizado y probabilizado del margen esperado, que es el objetivo marcado en este proceso de optimización para cada cliente, como combinación de las distintas opciones de prima propuesta y su efecto en la retención de las pólizas.

Los resultados del margen probable esperado para cada cliente se muestran a continuación:

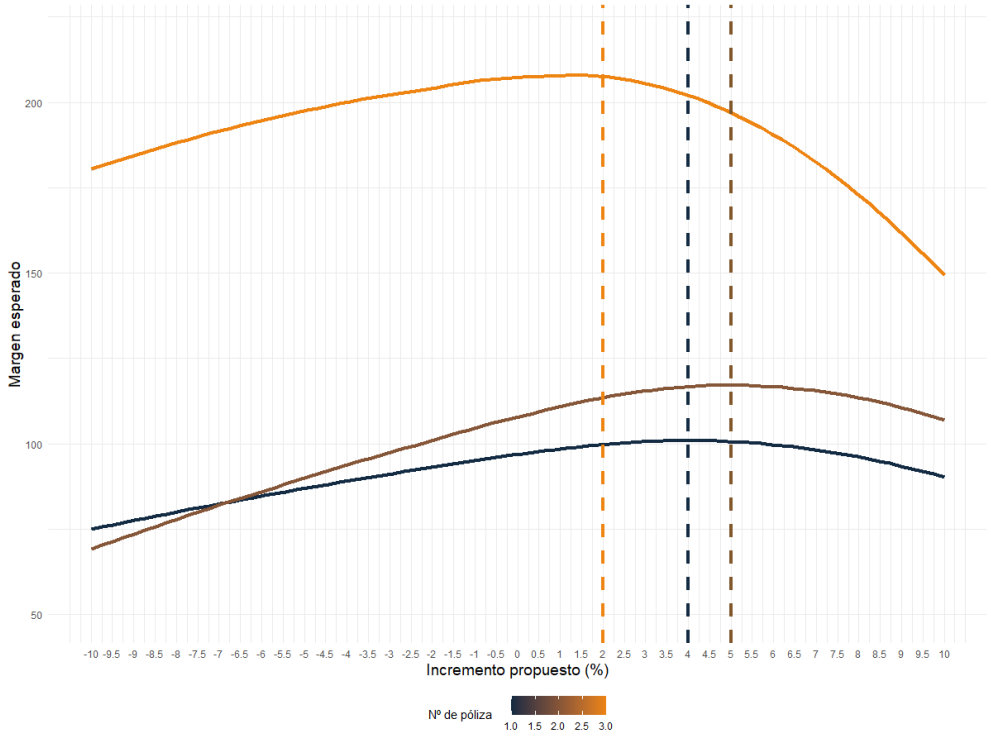


Figura 7.24: Margen probable esperado por cliente según la variación de prima propuesta

En el gráfico se observa como el máximo del margen individual para cada cliente se obtiene con una variación de prima distinta en cada póliza. Así, frente a una estrategia plana de subida de un 1%, los valores óptimos se alcanzan en estos clientes con variaciones de prima del 2%, el 4% y el 5%. Esto significa que si la entidad opta por una estrategia plana por póliza, (1%), y no considera las elasticidades precio/demanda individuales, está sufriendo una pérdida por el coste de oportunidad del margen no ganado que tendría en la estrategia optimizada.

Para ver el efecto agregado se muestran las variaciones de prima óptima para las 20.712 pólizas en la siguiente figura.

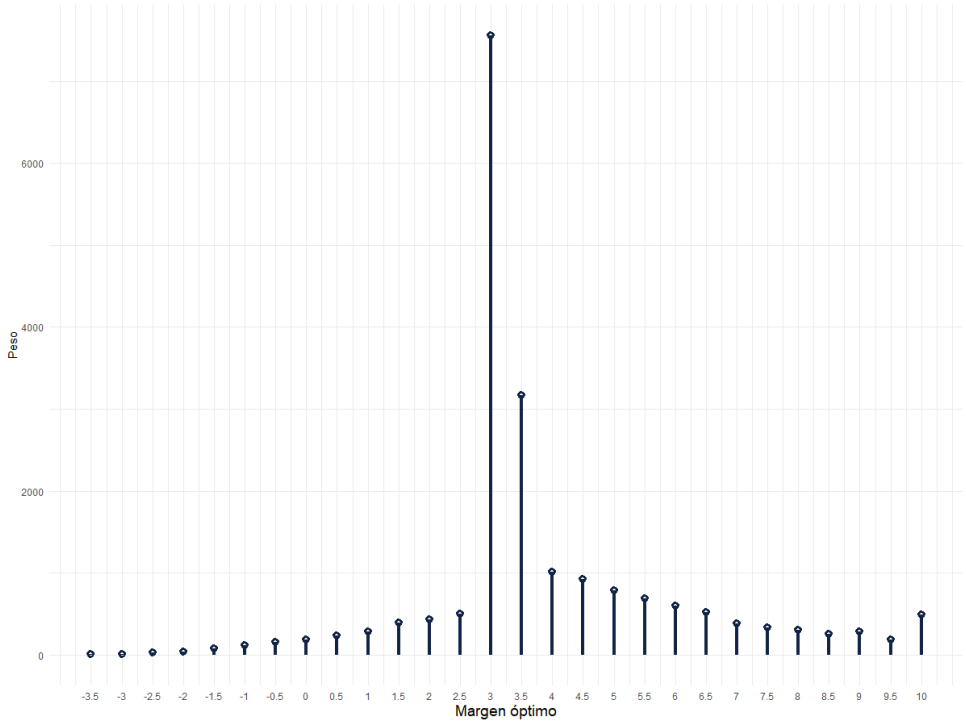


Figura 7.25: Histograma de variaciones óptimas de la renovación de la cartera

A la vista de los resultados obtenidos se aprecia como para esta cartera de pólizas, la estrategia óptima más repetida es la de una subida del 3%. Además, la mayoría de los clientes están dispuestos a pagar una prima por encima del 1% de la estrategia plana.

Si se cuantifica el margen según en cada alternativa propuesta se obtiene un beneficio técnico en la estrategia plana de 1,74 millones de euros. Si se optimiza el margen el valor se incrementa hasta los 1,94 millones de euros, lo que supone 201.000 euros más y que significa una mejora en el resultado de un 11,55%.

8. HERRAMIENTA DE TARIFICACIÓN

En esta sección se muestra el funcionamiento de una aplicación web creada íntegramente en el entorno de desarrollo R Shiny. El objetivo es construir una aplicación versátil que sea sencilla de implementar en los sistemas operativos de las entidades.

Se plantea como un complemento ante tareas como el *testing* de los precios en entorno de preproducción, la derivación del precio óptimo por póliza, el uso como tarificador para la estructura comercial y finalmente la monitorización de los KPIs esperados marcados por la dirección de la entidad.

Ahora mismo, existen dos softwares comerciales que ocupan este nicho de mercado en cuanto a optimización matemática de primas. Sin embargo, dado su elevado precio hace que muchas entidades más pequeñas no puedan disponer de esta tecnología. Herramientas como esta podrían suponer una alternativa para poder aplicar este tipo de metodologías.

A continuación, se adjuntan algunas salidas de la apariencia beta de la aplicación:

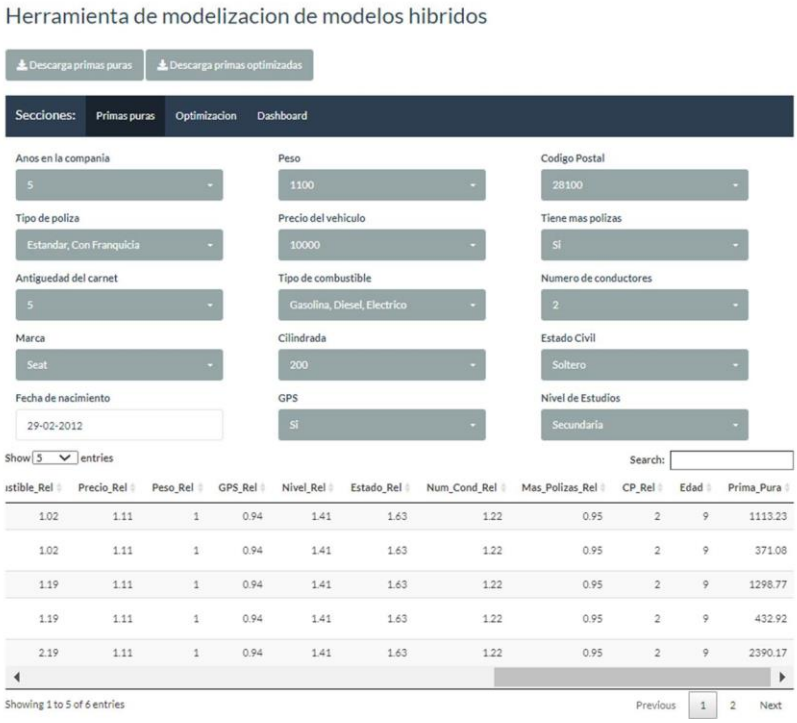


Figura 8.1: Datos de entrada para el cálculo de la prima pura

Figura 8.2: Optimización individual por póliza

9. CONCLUSIONES

En los últimos años las técnicas predictivas de Inteligencia Artificial se han introducido para fines muy diversos dentro de la sociedad. El ámbito asegurador no es ajeno a estos algoritmos. En el campo actuarial están empleándose estas nuevas técnicas para la tarificación de primas, el cálculo de las provisiones técnicas o los modelos de capital de solvencia.

En cuanto al cálculo de primas, los modelos de Inteligencia Artificial mejoran, en muchos casos, las capacidades predictivas de los modelos de tarificación tradicional (como los GLMs). No obstante, además de la mejora en las estimaciones antes de implantar una técnica concreta en los procesos de cálculo de prima, las entidades aseguradoras deben considerar otros elementos como los tiempos de implementación, la estabilidad de los cálculos y, sobre todo, la transparencia de los algoritmos hacia los distintos grupos de interés como supervisores o tomadores.

Adicionalmente, la falta de verificabilidad que presentan estas nuevas técnicas de Inteligencia Artificial pueden dificultar el control de los principios de no discriminación, equidad y suficiencia de las tarifas, aunque se mejore la capacidad predictiva de los modelos actuariales. Una alternativa que se presenta en este trabajo es el empleo de técnicas híbridas de tarificación, que combinan la mejora predictiva de los nuevos modelos de Inteligencia Artificial, con los algoritmos tradicionales que resultan más transparentes y fácilmente trasladables a los sistemas operacionales de emisión de las entidades aseguradoras.

Además, en los próximos años las características del vehículo cobrarán un mayor peso en la evaluación de la siniestralidad como consecuencia de las ayudas a la conducción que incorporan los coches (ADAS). Estas características reducirán previsiblemente el peso que tienen, en el cálculo de la prima, las variables específicas sobre los hábitos de conducción.

En cuanto al seguro de automóviles, las técnicas más sofisticadas de tarificación son aquellas que, por un lado, combinan distintas metodologías estadísticas para la mejora de los modelos predictivos y, por el otro, modelan estadísticamente no sólo

el riesgo asociado a la frecuencia y la severidad sino también el comportamiento del cliente frente a la prima ofrecida.

La técnica de tarificación que ofrece mejores resultados a las compañías aseguradoras es la optimización individual de primas, ya que permite maximizar el margen técnico y la retención (o conversión) de pólizas para la entidad aseguradora. En este trabajo, para un ejemplo de renovación de cartera del seguro de automóviles, se ha comprobado como esta técnica supera los resultados obtenidos a través de otras estrategias de tarificación tradicional que no tienen en cuenta los modelos de demanda de los clientes.

10. FUTURAS LÍNEAS DE INVESTIGACIÓN

Este trabajo abre algunas líneas de investigación sobre las que me gustaría profundizar. En primer lugar, la evolución de los avances en los tipos de combustión o las ayudas a la conducción de los vehículos, van a provocar un cambio en las variables determinantes del comportamiento del riesgo en el seguro de automóviles. En la actualidad, la información disponible sobre esas características es escasa y poco granular. Una de las líneas de trabajo es la medición del efecto que tienen esos atributos del automóvil en el comportamiento de la siniestralidad para enriquecer el *scoring* de vehículo con esta nueva información.

Adicionalmente me gustaría ampliar el abanico de técnicas de Inteligencia Artificial que se podrían utilizar en la creación de modelos híbridos, en su aplicación a la tarificación, por ejemplo, con la utilización de otros algoritmos como *XGBoost*, *Light GBM*, *CatBoost*, *Ridge*, *Lasso* y redes neuronales entre otras para mejorar la capacidad predictiva, la búsqueda de interacciones complejas o la reducción de dimensionalidad de los modelos.

En el campo de la optimización matemática tengo intención de profundizar en la derivación de modelos de optimización individual. En esa línea, pretendo desarrollar la optimización de factores de riesgo y la fijación de restricciones globales para la

herramienta, que hace más compleja la expresión de la función a optimizar y plantea algunos problemas computacionales.

Además de la consideración del margen o la retención, también se podrían tener en cuenta otras variables a la hora de realizar el proceso de optimización matemática de la prima. Una de ellas podría ser la minimización del capital de solvencia requerido por póliza, especialmente si la entidad dispone de un modelo interno.

Finalmente, voy a seguir trabajando en la implementación operativa de todas estas técnicas a partir de la utilización de software libre, añadiendo las mejoras anteriormente expresadas en la herramienta creada en R Shiny e incorporando nuevas funcionalidades como la simulación de medición de impactos o la optimización de la nueva producción.

11. BIBLIOGRAFÍA

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. y Thandi, N. (2007). *A practitioner´s guide to generalized linear models*. Towers Watson.

https://www.aktuarai.lt/wpontent/uploads/2018/06/Anderson_et_al_Edition_3.pdf

Amraoui, A., Bishop, D., Kroetz, P., Lorenz, J.T., Martos, C. y Sancier, S. (2009). *The power of going direct*. Mckinsey's Insurance Practice.

Casualty Actuarial and Statistical Task Force. (2015). *Price Optimization White Paper*.
https://www.naic.org/documents/committees_c_catf_related_price_optimization_white_paper.pdf.

Coskun, S. (2016*). Introducing credibility theory into GLMs for ratemaking on auto portfolio*. Centre d´etudes actuarielles.

Cummings, D. (2015). *Optimization applications in insurance*. [Presentación de PowerPoint].

De Jong, P., y Heller, G.Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

DIRECCIÓN GENERAL DE TRAFICO. (2021). *Series históricas del parque de vehículos*.

https://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/parque-vehiculos/series-historicas/.

Diana, A., Griffin, J.E., Oberoi, J. y Yao, J. (2018). *Machine learning methods for insurance applications*. Society of Actuaries.

https://www.soa.org/resources/research-reports/2019/machine-learning-methods/

EARNIX (2020). *Optimization in Earnix*.

Evans, E., Hughes, C. (2019). *Risk Pricing with XGBoost.* [Presentación de PowerPoint].

https://www.youtube.com/watch?v=sOyMLB1SsFk

Feldblum, S. y Brosius, J.E. (2002). *The minimum bias procedure. A practitioner´s guide*. Casualty Actuarial Society.

https://www.casact.org/sites/default/files/database/forum_02fforum_02ff591.pdf

Fernández, D. (19 de agosto, 2016). Historia y leyenda del primer seguro de coche. *El Economista*.

https://www.eleconomista.es/ecomotor/motor/noticias/7770897/08/16/Historia-y-leyenda-del-primer-seguro-de-coche.html.

Feldman, J., Brown, R. (2005). *Risk and Insurance.* Education and examination committee of the society of actuaries.

García Moreno, P. (23 de febrero, 2021). Los coches híbridos sufren menos accidentes de tráfico. *Cinco Días.*

https://cincodias.elpais.com/cincodias/2021/02/23/companias/1614084794_107146.html

Garrido, J. y Zhou, J. (2009). Full credibility with generalized linear and mixed models. *Astin Bulletin*. Vol. 39, nº 1, pp. 61-80.

Guillén, M. y Pesantez-Narvaez, J. (2018). Machine learning y modelización predictiva para la tarificación en el seguro de automóviles. *Anales del Instituto de Actuarios Españoles*, 4ª época, 24, pp. 123-147.

https://www.actuarios.org/wp-content/uploads/2018/11/123_147_A06.pdf

Hartwig, R.P. (2015). *Price optimization in auto insurance markets*. National Conference of Insurance Legislators.

https://www.iii.org/presentation/price-optimization-in-auto-insurance-markets-actuarial-economic-and-regulatory-considerations-071715

Hastie, T. y Tibshirani, R. (1986). Generalized additive models. *Statistical Science*. Vol. 1, nº 3, pp. 297-318.

Investigación Cooperativa Entidades Aseguradoras. (2018). *Impacto del coche del futuro en el ramo de autos*. Número 281.

Investigación Cooperativa Entidades Aseguradoras. (2021). *El seguro de automóviles a diciembre. Año 2020*. Número 1.642.

Johnston, L. (2020). *The evolution of the motor insurance industry*. [Thesis, University of Essex].

Kaivanipour, K. (2015). *Non-life insurance pricing using the Generalized Additive Model, Smoothing Splines and L-curves*. Royal Institute of Technology. Stockholm.

https://www.diva-portal.org/smash/get/diva2:818695/FULLTEXT01.pdf

MAPFRE Economics. (2020). *El mercado español de seguros en 2019.* Fundación MAPFRE.

Maréchal, X. (2020). *Machine Learning Applications to Non-Life Pricing* [Presentación de PowerPoint].

Mildenhall, S. (1999). *A systematic relationship between minimum bias and generalized linear models*. Casualty Actuarial Society.

https://www.casact.org/sites/default/files/database/proceed_proceed99_99317.pdf

McCullagh, P. y Nelder, J.A. (1989). *Generalized linear models*. Chapman and Hall.

Nelder, J.A. y Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*. Vol. 135, nº 3, pp. 370-384.

Ohio Department of Insurance. (2015). *Price optimization*.

https://us.eversheds-sutherland.com/portalresource/OHBulletin2015-01.pdf

Panlilio, A., Canagaretna, B., Perkins, S., Preez, V. y Zhixin, L. (2018). *Practical application of machine learning within actuarial work*. Institute and Faculty of Actuaries.

https://www.actuaries.org.uk/system/files/field/document/Practical%20Application%20of%20Machine%20Learning%20within%20Actuarial%20Work%20Final%20%282%29_feb_2018.pdf

Rodriguez-Pardo, J.M. (2017). El actuario ante Insurtech. *Revista Actuarios*, 41, 12-16.

https://app.mapfre.com/documentacion/publico/es/catalogo_imagenes/grupo.do?path=1095318

Santoni, A., Gómez, F. (2007). *Sophisticated price optimization methods*. [Presentación de PowerPoint].

https://www.casact.org/sites/default/files/presentation/ratesem_2008_handouts_gomez.pdf

Tiwari, A. (30 de Marzo, 2020). Modeling Insurance Claim Severity. *Medium.*

https://medium.com/swlh/modeling-insurance-claim-severity-b449ac426c23

Werner, G. y Modlin C. (2016). *Basic Ratemaking*. Casualty Actuarial Society.

https://www.casact.org/sites/default/files/old/studynotes_werner_modlin_ratemaking.pdf

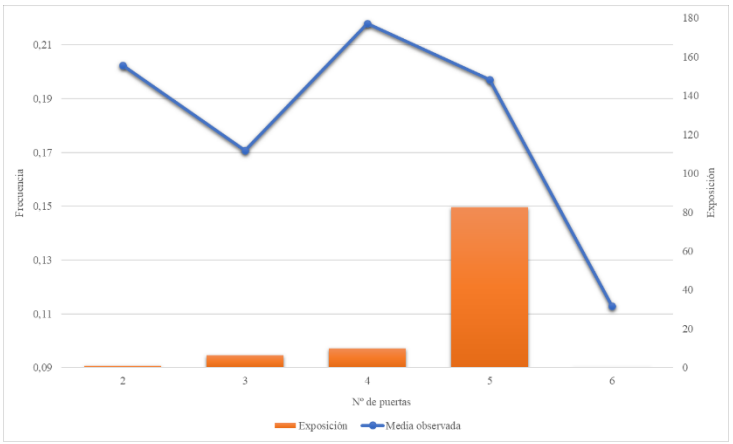
Witten, I.H., Frank E., Hall, M.A. y Pal, C.J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

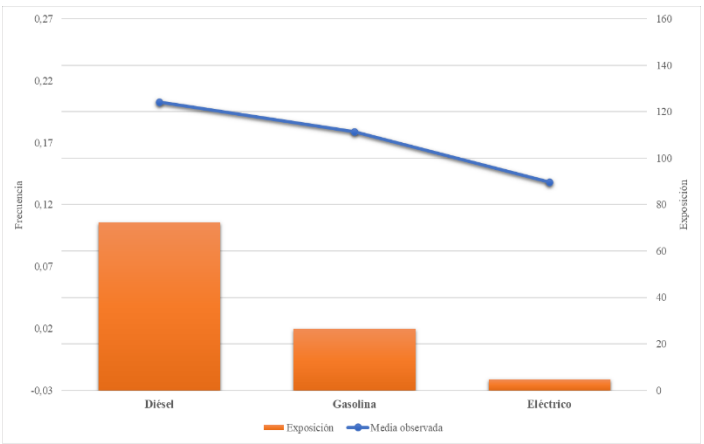
Zhou, J., Deng, D. (2019). *GLM vs. Machine Learning with Case Studies in Pricing* [Presentación de PowerPoint].

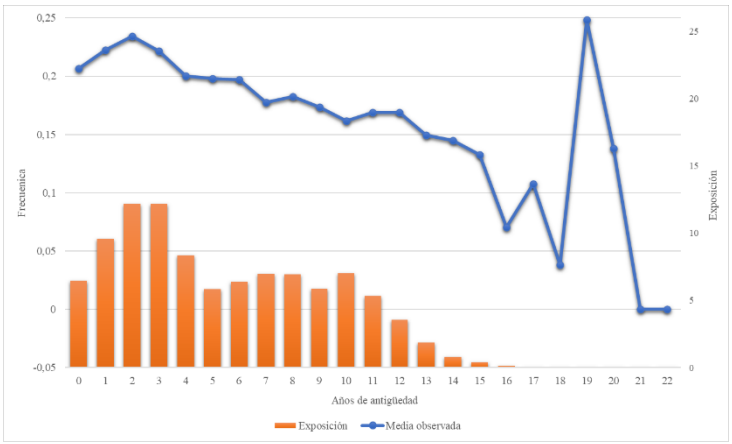
https://www.casact.org/sites/default/files/presentation/annual_2019_presentations_c-22_zhou.pdf

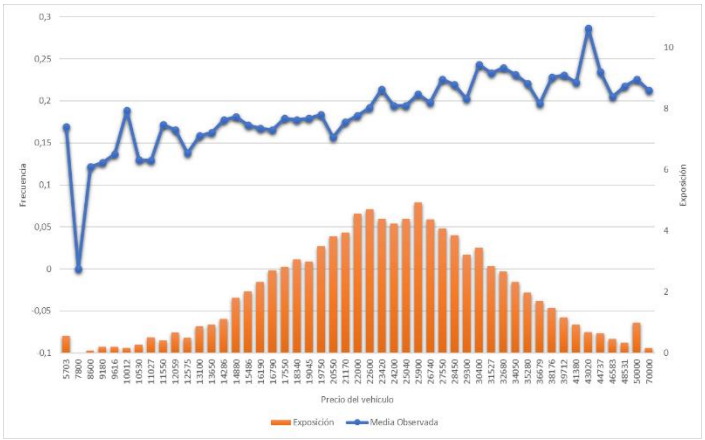
12. ANEXOS

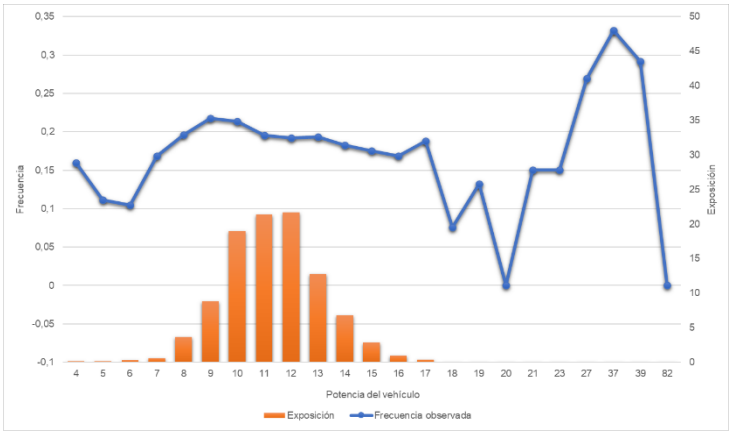
12.1 Análisis univariable de las variables de vehículo para el modelo de scoring

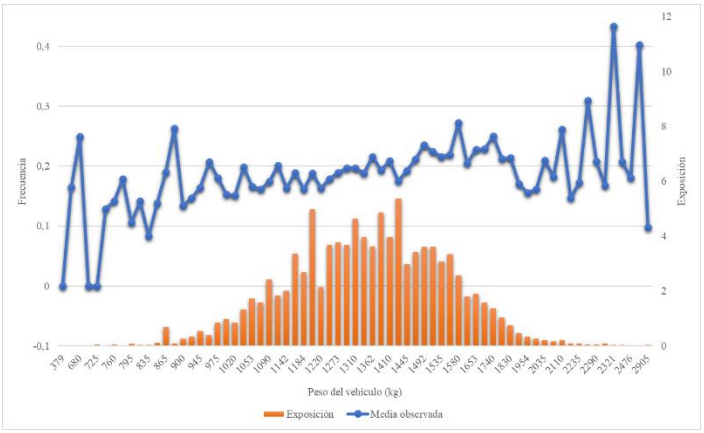


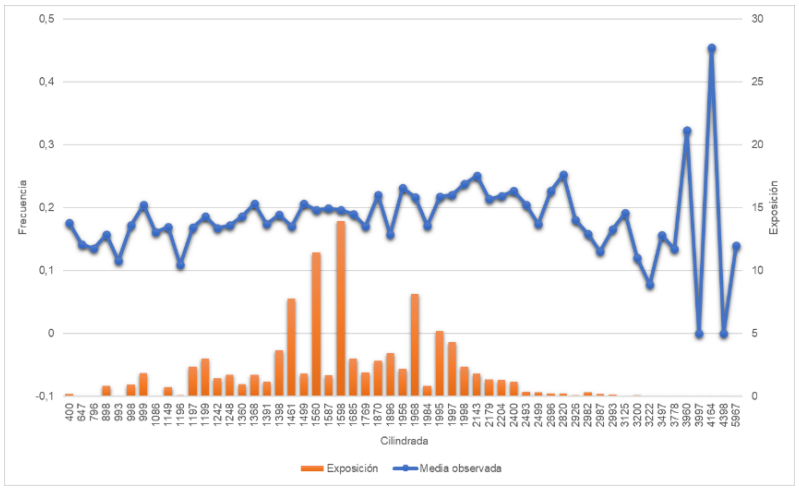


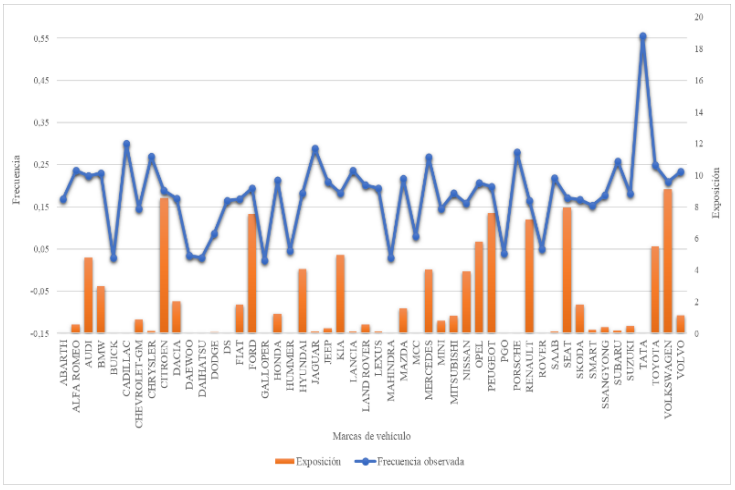


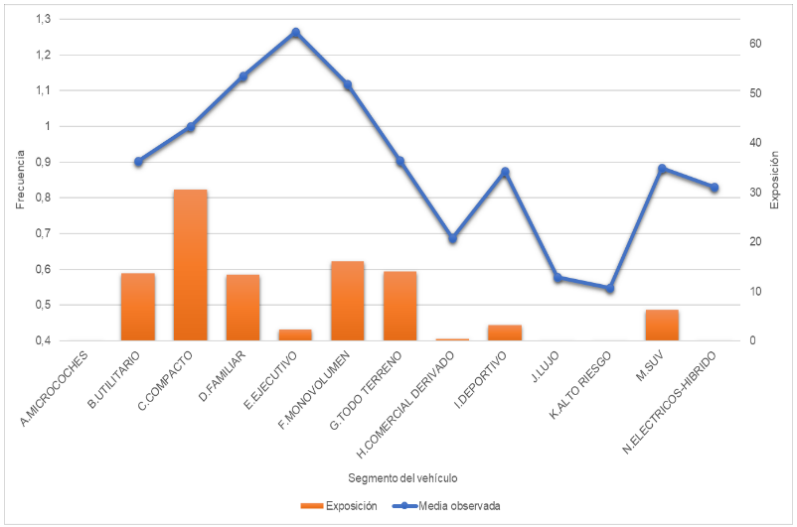


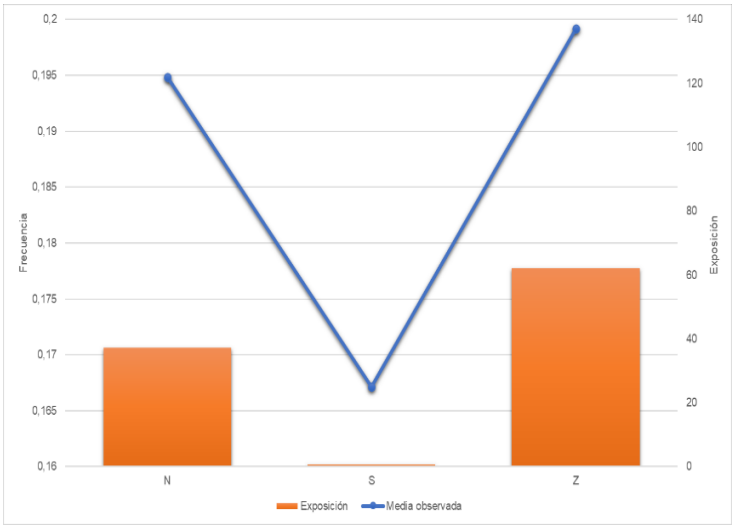




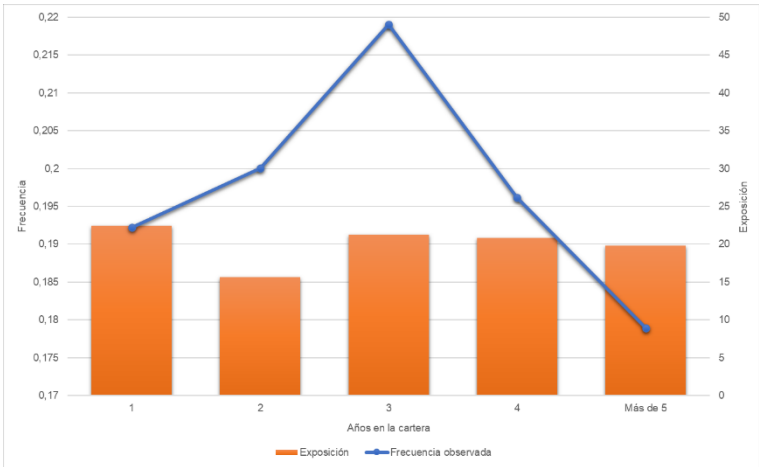

Frecuencia
Exposición
Cilindrada
Exposición
Media observada

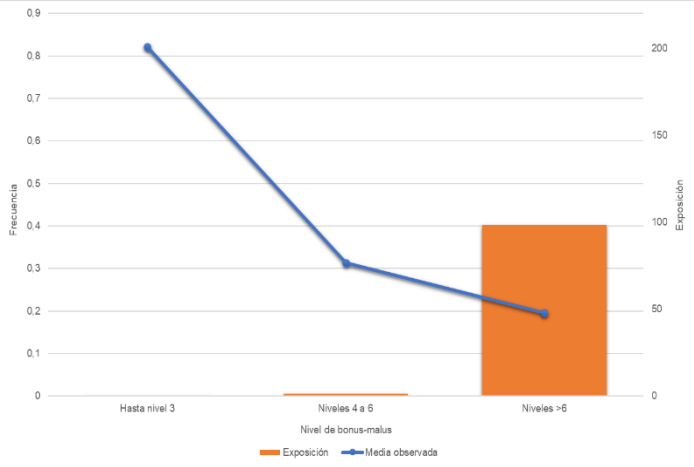

Frecuencia
Exposición
Marcas de vehículo
Exposición
Frecuencia observada

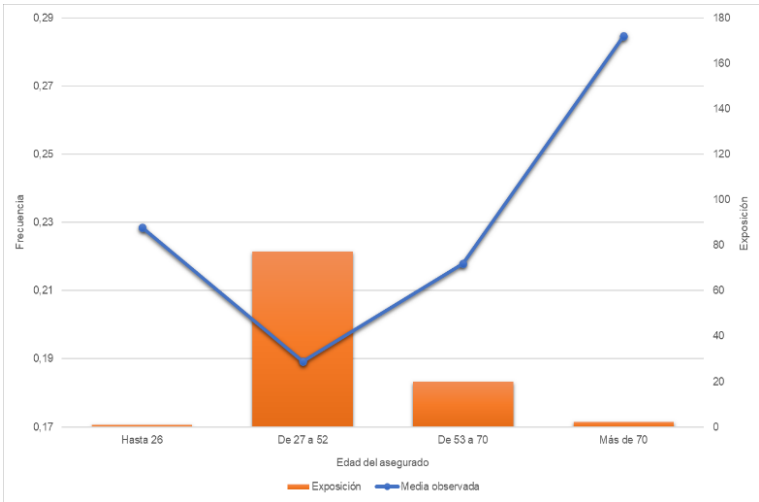

Frecuencia
Exposición
A.MICROCOCHES
B.UTILITARIO
C.COMPACTO
D.FAMILIAR
E.EJECUTIVO
F.MONOVOLUMEN
G.TODO TERRENO
H.COMERCIAL DERIVADO
I.DEPORTIVO
J.LUJO
K.ALTO RIESGO
M.SUV
N.ELECTRICOS-HIBRIDO
Segmento del vehículo
Exposición
Media observada

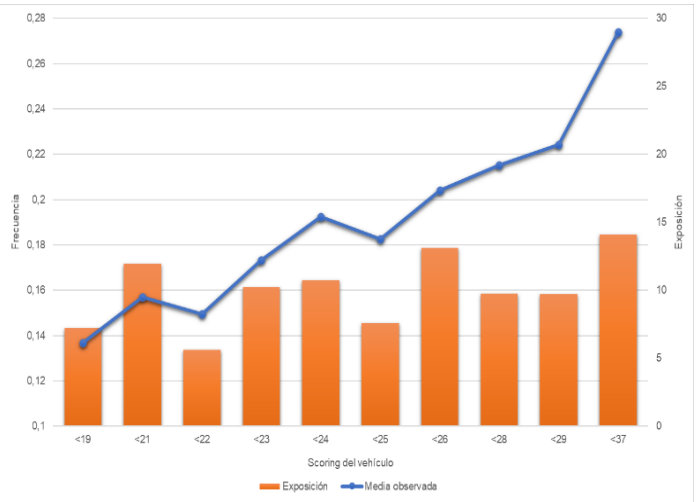

Frecuencia
Exposición
N
S
Z
Exposición
Media observada

12.2 Análisis univariable del modelo de frecuencia

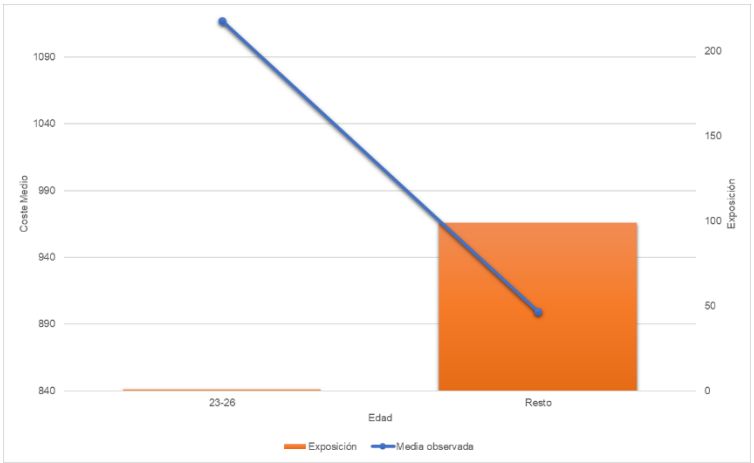


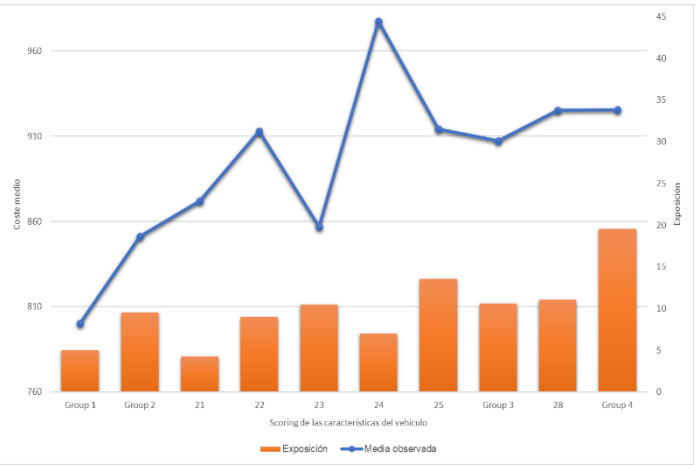


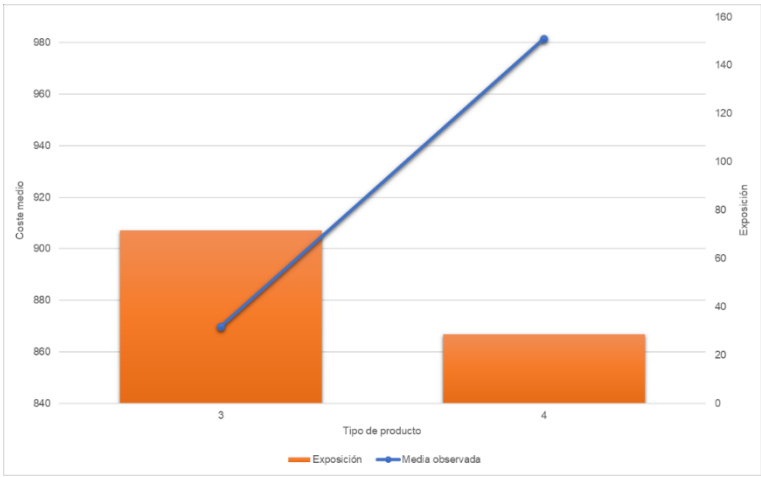




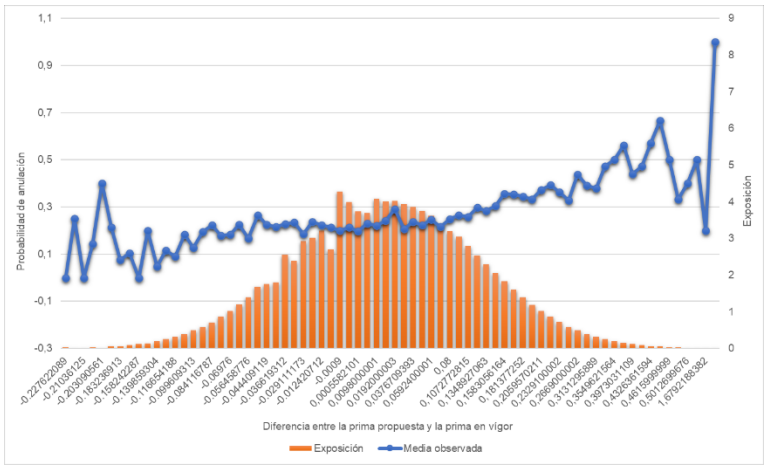
12.3 Análisis univariable del modelo de severidad

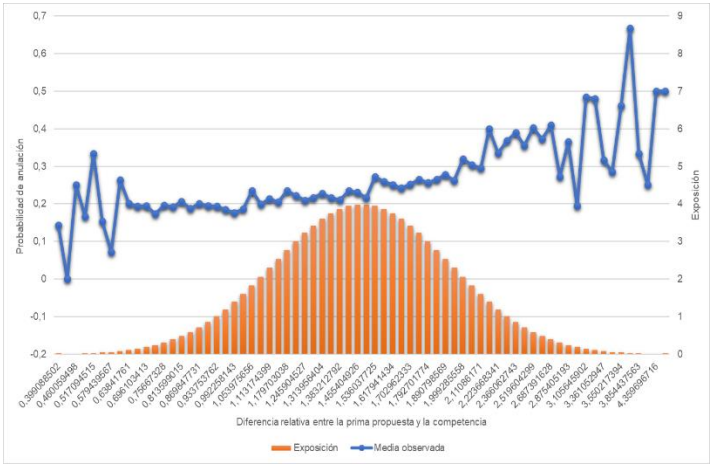


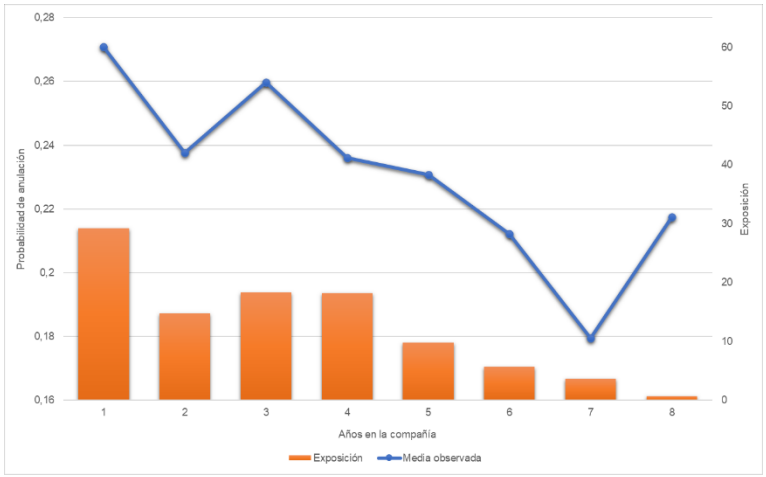


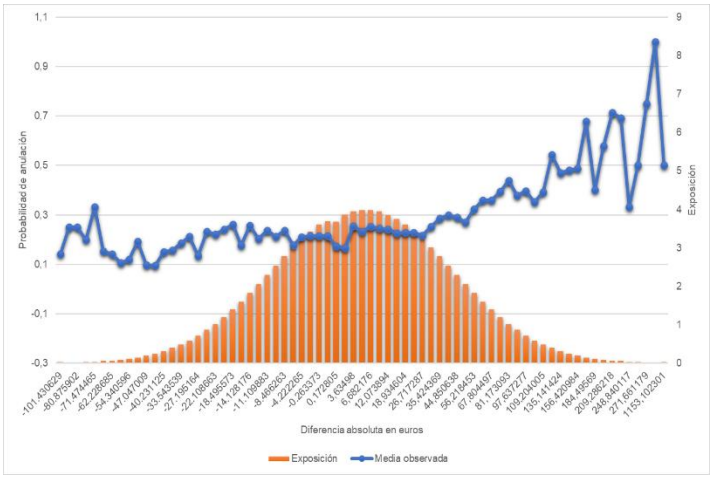


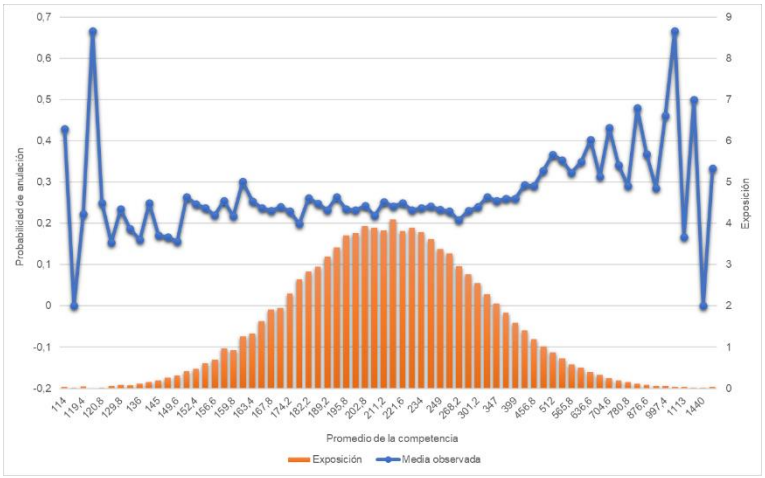
12.4 Análisis univariable del modelo de anulación











12.5 Código

```
################################################################################

                    CAR GROUP – Daños personales

################################################################################


library(tidyverse)

library(knitr)

library(gtools)

library(esquisse)


datos= read.csv2("C:/Users/juanc/OneDrive/Escritorio/TFM/Secciones/4_Los datos/BBDD_tfm.csv")


datos <- datos %>% filter(MODAL>2)

data_car_group<-
datos[,c("EVYTOT","DP_INC","NUMDP","BONOREN","CARAGE","CARCC","CARCV","CARDOOR
S","CARFUEL","CARGPS","CARKW","CARMAKE","CARMODEL","CARPPOT","CARPVP","CARSE
GM","CARTAR","EQUIFAX","MAINAGE","MAINDLY","MAINDSEX","MAINPROV","MODAL","PAYF
REQ","POLCHAN","RNLYEAR")]


EQUIFAX_LEVELS <- c(1,2,3,4,5,6)


################################################################################

                        LIMPIEZA DE LA BBDD

################################################################################

datos_limpios <- data_car_group %>%

            filter(EVYTOT>0 ,

                DP_INC >= 0,

                CARPVP > 537,

                MAINAGE - MAINDLY > 18,
```

```
                    CARDOORS > 0)  %>%


                mutate(EVYTOT = ifelse(EVYTOT > 1, 1 ,EVYTOT),
                    FREQ = NUMDP / EVYTOT,

CARFUEL=ifelse(CARFUEL=="D","Diesel",ifelse(CARFUEL=="G","Gasolina","Electrico")),
                    CARGPS = ifelse(CARGPS=="Z","N",CARGPS),
                    EQUIFAX = ifelse(EQUIFAX %in% EQUIFAX_LEVELS,"Z",EQUIFAX)) %>%


                mutate_if(is.character, as.factor) %>%
                mutate_if(is.numeric, list(~ replace(., is.na(.), median(., na.rm = TRUE))))


datos_limpios[is.na(datos_limpios)] <- mode(datos_limpios)


datos_coches<-
datos_limpios[,c("CARAGE","CARCC","CARCV","CARDOORS","CARFUEL","CARGPS","CARKW",
"CARMAKE","CARMODEL","CARPPOT","CARPVP","CARSEGM","CARTAR")]


################################################################################
                        EXPLORACIÓN DE LAS CORRELACIONES
################################################################################
datos_numericos<-
datos_limpios[,c("CARAGE","CARCC","CARCV","CARDOORS","CARKW","CARPPOT","CARPVP",
"CARTAR")]
corr <- round(cor(datos_numericos), 2)
corr
```

```
################################################################################

                                AGRUPACIONES

################################################################################


datos_trameados <- datos_limpios %>%


                    mutate(CARAGE = as.factor(ifelse(CARAGE >= 15, ">15", CARAGE)),

                           CARPPOT                                                    =
as.factor(ifelse(CARPPOT<8,"<8",ifelse(CARPPOT>16,">16",round(CARPPOT,0)))),

                           CARKW                                                      =
as.factor(ifelse(CARKW<46,"<46",ifelse(CARKW>200,">200",round(CARKW,0)))),

                           CARCV                                                      =
as.factor(ifelse(CARCV<63,"<63",ifelse(CARCV>200,">200",round(CARCV,0)))),

                           CARTAR                                                     =
as.factor(ifelse(CARTAR<900,"<900",ifelse(CARTAR>2110,">2110",round(CARTAR,0)))),

                           CARCC                                                      =
as.factor(ifelse(CARCC<1200,"<1200",ifelse(CARCC>2400,">2400",round(CARCC,0)))),

                           MAINAGE = as.factor(ifelse(MAINAGE > 75, ">75", MAINAGE)),

                           MAINDLY = as.factor(ifelse(MAINDLY > 50, ">50", MAINDLY)))


quantiles_5_precio = quantile(datos_trameados$CARPVP,seq(0,1,0.02))


Cuantiles_5_precio= cut(datos_trameados$CARPVP,quantiles_5_precio,right = FALSE,dig.lab=10)


kable(table(Cuantiles_5_precio),caption = "Distribución según el precio",digits = 2, align =
"c",col.names = c("Niveles","Frecuencia"),format.args = list(big.mark = ","))


datos_trameados <- datos_trameados %>%

                    mutate(CARPVP= cut(CARPVP,quantiles_5_precio,right = FALSE,dig.lab=10))
```

```
################################################################################

                    AGRUPACIÓN MEDIANTE UN CART

################################################################################

library(cluster)

library(Rtsne)

library(rattle)


df_cluster <- datos_trameados[,c("FREQ","CARMAKE")]

arbol_freq <- rpart(formula = FREQ ~ .,data = df_cluster,method="anova",control = rpart.control(cp = 0.00001))

rpart.plot(arbol_freq,tweak = 0.8)

fancyRpartPlot(arbol_freq,tweak = 0.6,type=1)


################################################################################

                    CLUSTERING USANDO LA DISTANCIA DE GOWEN

################################################################################


df_cluster<- df_cluster[sample(nrow(df_cluster),5000),]


gower_dist <- daisy(df_cluster, metric = "gower")

gower_mat <- as.matrix(gower_dist)


sil_width <- c(NA)

for(i in 2:30){

  pam_fit <- pam(gower_dist, diss = TRUE, k = i)

  sil_width[i] <- pam_fit$silinfo$avg.width

}
```

```
Data_error <- data.frame(Num=1:30,Error=sil_width)


ggplot(Data_error) +
  aes(x = Num, y = Error) +
  geom_point(
    shape = "circle",
    size = 3.35,
    colour = "#228B22"
  ) +
  geom_smooth(span = 0.19) +
  labs(
    x = "Número de clusters",
    y = "Score Silhoette",
    title = "Método Silhoutte"
  ) +
  theme_bw()


k <- 3
pam_fit <- pam (gower_dist, diss = TRUE, k)
pam_results <- df_cluster%>%
  mutate (cluster = pam_fit $ clustering)%>%
  group_by (cluster)%>%
  do (the_summary = summary (.) )


pam_results $ the_summary
tsne_obj <- Rtsne (gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj $ Y%>%
  data.frame ()%>%
```

```
  setNames (c ("X", "Y"))%>%
  mutate (cluster = factor (pam_fit $ clustering))
ggplot (aes (x = X, y = Y), data = tsne_data) +
  geom_point (aes (color = cluster))+theme_apa()
pam_results $ the_summary


################################################################################
                                SCORING MEDIANTE GLM
################################################################################


Modelo scoring GLM
modelo_glm = glm(NUMDP ~ CARAGE + CARGPS,
            family = poisson(link="log"),
            data = BBDD_dp,offset = log(EVYTOT))
BBDD_dp_xgb$RELATIVITY = ifelse(BBDD_dp_xgb$CARAGE==0,Relatividades[1],
                        ifelse(BBDD_dp_xgb$CARAGE==1,Relatividades[2],
                            ifelse(BBDD_dp_xgb$CARAGE==2,Relatividades[3],
                                ifelse(BBDD_dp_xgb$CARAGE==3,Relatividades[4],
                                    ifelse(BBDD_dp_xgb$CARAGE==4,Relatividades[5],
                                        ifelse(BBDD_dp_xgb$CARAGE==5,Relatividades[6],
                                            ifelse(BBDD_dp_xgb$CARAGE==6,Relatividades[7],
                                                ifelse(BBDD_dp_xgb$CARAGE==7,Relatividades[8],

ifelse(BBDD_dp_xgb$CARAGE==8,Relatividades[9],
ifelse(BBDD_dp_xgb$CARAGE==9,Relatividades[10],
ifelse(BBDD_dp_xgb$CARAGE==10,Relatividades[11],
ifelse(BBDD_dp_xgb$CARAGE==11,Relatividades[12],
ifelse(BBDD_dp_xgb$CARAGE==12,Relatividades[13],
ifelse(BBDD_dp_xgb$CARAGE==13,Relatividades[14],
ifelse(BBDD_dp_xgb$CARAGE==14,Relatividades[15],
```

```
ifelse(BBDD_dp_xgb$CARAGE==15,Relatividades[16],
ifelse(BBDD_dp_xgb$CARAGE==16,Relatividades[17],
ifelse(BBDD_dp_xgb$CARAGE==17,Relatividades[18],
ifelse(BBDD_dp_xgb$CARAGE==18,Relatividades[19],
ifelse(BBDD_dp_xgb$CARAGE==19,Relatividades[20],
ifelse(BBDD_dp_xgb$CARAGE==20,Relatividades[21],
ifelse(BBDD_dp_xgb$CARAGE==21,Relatividades[22],Relatividades[23])))))))))))))))))))))))



################################################################################

                                SCORING MEDIANTE GBM

################################################################################


data_agreg <- data.frame(data_agreg,conteo=c(1:17609))

ggplot(data_agreg, aes(x=conteo)) +

  geom_point(aes(y = FREQ), color = "darkred") +

  geom_line(aes(y = predicciones_freq), color="steelblue", linetype="twodash")


caret::RMSE(data_agreg$predicciones_freq, data_agreg$FREQ)

caret::MAE(data_agreg$predicciones_freq, data_agreg$FREQ)

table(XGB_DF$FREQ)


gbm.fit <- gbm(

  formula = FREQ  ~ .,

  distribution = "gaussian",

  data = train_set,

  n.trees = 10000,

  interaction.depth = 1,

  shrinkage = 0.001,

  cv.folds = 5,
```

```
  n.cores = NULL,
  verbose = FALSE
)
pred_glm_score <- read.csv2("C:/Users/juanc/OneDrive/Escritorio/TFM/Secciones/5_Algoritmos a utilizar/CAR GROUPS/RMSE_MAE_GLM_SCORING.csv")

caret::RMSE(pred_glm_score$Freq.Estimada, pred_glm_score$Freq.Observada)
caret::MAE(pred_glm_score$Freq.Estimada, pred_glm_score$Freq.Observada)

predicciones_RMSE = data.frame(pred_glm_score=pred_glm_score,predicciones_GBM=predicciones_freq)

predicciones_RMSE <- predicciones_RMSE %>%
                    mutate(Residuos_glm=pred_glm_score.Freq.Observada-pred_glm_score.Freq.Estimada,
                           Residuos_gbm=pred_glm_score.Freq.Observada-predicciones_GBM,
                           Datos = 1:length(predicciones_RMSE$predicciones_GBM))
ggplot(predicciones_RMSE, aes(x=Datos)) +
  geom_point(aes(y = Residuos_gbm), color="blue") +
  geom_point(aes(y = Residuos_glm), color = "red")

write.csv(predicciones_RMSE,"C:/Users/juanc/OneDrive/Escritorio/TFM/Secciones/5_Algoritmos a utilizar/CAR GROUPS/File Name.csv", row.names = FALSE)

datos_glm_freq <- datos_trameados[,4:27]

OTHER <- c("MAHINDRA","PORSCHE","CADILLAC","DAEWOO","DAIHATSU","GALLOPER","TATA","ROVER")
```

```
datos_glm_freq <- datos_glm_freq %>%
        mutate(CARMAKE=ifelse(CARMAKE == "MAHINDRA","OTROS",
                        ifelse(CARMAKE =="PORSCHE","OTROS",
                            ifelse(CARMAKE ==" CADILLAC","OTROS",
                                ifelse(CARMAKE =="DAEWOO","OTROS",
                                    ifelse(CARMAKE =="DAIHATSU","OTROS",
                                        ifelse(CARMAKE =="GALLOPER","OTROS",
                                            ifelse(CARMAKE =="TATA","OTROS",
                                                ifelse(CARMAKE
=="ROVER","OTROS",CARMAKE)))))))))

#############################################################################
                                    MODELO FRECUENCIA
#############################################################################
datos_modelo_brutos<-       read_excel("C:/Users/juanc/OneDrive/Escritorio/TFM/Secciones/4_Los
datos/BBDD_TFM_SCORING.xlsx")

datos_modelo_brutos_1                                                                    <-
datos_modelo_brutos[,c("EVYTOT","DP_INC","NUMDP","BONOREN","MAINAGE","MAINDLY","M
ODAL","RNLYEAR","Scoring")]

datos_modelo <- datos_modelo_brutos_1 %>%
                filter(EVYTOT > 0, DP_INC >= 0)  %>%
                mutate(EVYTOT = ifelse(EVYTOT > 1, 1 ,EVYTOT))

datos_modelo_freq <- datos_modelo %>%

 mutate(MODAL=factor(MODAL),
```

```
      Scoring=ifelse(Scoring<=19,"<19",
                     ifelse(Scoring<=21,"[20,21]",
                            ifelse(Scoring<=22,"[22]",
                                   ifelse(Scoring<=23,"[23]",
                                          ifelse(Scoring<=24,"[24]",
                                                 ifelse(Scoring<=25,"[25]",
                                                        ifelse(Scoring<=26,"[26]",
                                                               ifelse(Scoring<=28,"[27,28]",
                                                                      ifelse(Scoring<=29,"[29]",">29")))))))),
      RNLYEAR=ifelse(RNLYEAR>5,">5",RNLYEAR),


      BONOREN=ifelse(BONOREN<4,"<4",
                     ifelse(BONOREN<6,"[4-6]",">=7")),


      MAINAGE=ifelse(MAINAGE<27,"<27",
                     ifelse(MAINAGE<52,"[28-52]",
                            ifelse(MAINAGE<71,"[53-70]",">=71"))))


 modelo_frecuencia = glm(NUMDP ~ MAINAGE + MODAL + Scoring + RNLYEAR + BONOREN,
                     family = poisson(link="log"),
                     data = datos_modelo_freq,offset = log(EVYTOT))


Relatividades_Frecuencia <- exp(coef(modelo_frecuencia))
```

```
################################################################################

                                MODELO SEVERIDAD

################################################################################
datos_modelo_sev <- datos_modelo_freq %>%
                        filter(NUMDP>0,DP_INC>0)


modelo_severidad = glm(DP_INC ~ MAINAGE+MODAL ,
                               family = Gamma(link="log"),
                               data = datos_modelo_sev,offset = log(NUMDP))
Relatividades_Severidad <- exp(coef(modelo_severidad))


path_out = 'C:/Users/juanc/OneDrive/Escritorio/TFM/Secciones/5_Algoritmos a utilizar'
write.csv(Relatividades_Frecuencia,fileName_1)
write.csv(Relatividades_Severidad,fileName_2)
head(datos_modelo_freq)
datos_modelo_freq <- datos_modelo_freq %>%
                        mutate(contador=1:length(datos_modelo_freq))


datos_modelo_sev <- datos_modelo_sev %>%
 mutate(contador=1:9087)
 )
datos_modelo_sev %>%
 filter(DP_INC >= 0L & DP_INC <= 10000L) %>%
 ggplot() +
 aes(x = DP_INC) +
 geom_density(adjust = 1L, fill = "#112446") +
 labs(x = "Coste del siniestro", y = "Densidad") +
 theme_minimal() +
```

```
 theme(
   axis.title.y = element_text(size = 14L),
   axis.title.x = element_text(size = 14L)
 )


##############################################################################
                                    PRIMA PURA
##############################################################################
datos_con_prima <- datos_modelo_sev %>%
                mutate(FREQ = exp(predict(modelo_frecuencia,datos_modelo_sev)),
                    SEV = exp(predict(modelo_severidad,datos_modelo_sev)))


datos_con_prima <- datos_con_prima %>%
                  mutate(PRIMA=FREQ*SEV)


datos_con_prima_1 <- datos_con_prima %>%
                     mutate(MODAL=ifelse(MODAL==3,"Daños propios con franquicia","Daños
propios"))


##############################################################################
                                RESIDUOS DEVIANZA
##############################################################################
resid_freq = rstandard(modelo_frecuencia, type='deviance')
resid_sev = rstandard(modelo_severidad, type='deviance')


ggplot(resid_freq) +
 aes(x = Valores) +
 geom_density(adjust = 1L, fill = "#112446") +
```

```
 labs(x = "Valores de los residuos", y = "Densidad") +
 theme_minimal() +
 theme(
   axis.title.y = element_text(size = 14L),
   axis.title.x = element_text(size = 14L)
 )
ggplot(resid_sev) +
 aes(x = Valores) +
 geom_density(adjust = 1L, fill = "#112446") +
 labs(x = "Valores de los residuos", y = "Densidad") +
 theme_minimal() +
 theme(
   axis.title.y = element_text(size = 14L),
   axis.title.x = element_text(size = 14L)
 )
##############################################################################
                                    OPTIMIZACIÓN
##############################################################################
numero_var <- seq(1,41,1)
numero_var
 i=c()
for (j in 1:4){
      for (i in numero_var){
             x[j,paste("variable",numero_var,sep="_")] = 0
           }
 }
 price_var <- seq(-.1,.1,0.005)
```

```
for (j in 1:4){
   for (i in numero_var){
      x[j,i+2] = (x[j,1]*price_var[i])*(x[j,2])*demanda[j,i]
    }
}
 precios = data.frame()


 for (j in 1:4){
          for (i in numero_var){
                precios[i,j]=x[j,i]
              }
 }
precios$contador = 1:41
data_ggp <- data.frame(x = precios$contador,
                    y = c(precios$Escenario_1, precios$Escenario_2, precios$Escenario_3),
                    group = c(rep("y1", nrow(precios)),
                              rep("y2", nrow(precios)),
                              rep("y3", nrow(precios))))
 gplot(data_ggp) +
 aes(x = x, y = y, colour = group) +
 geom_line(size = 1.1) +
 scale_color_hue(direction = 1) +
 labs(
   x = "Movimiento del precio",
   y = "Valor esperado",
   color = "Escenarios"
 ) +
 theme_minimal() +
```

```
  theme(
    legend.position = "bottom",
    axis.title.y = element_text(size = 14L),
    axis.title.x = element_text(size = 14L)
  )
graph_margen_total <- rbind(graph_margen,graph_margen2,graph_margen3)


ggplot(graph_margen_total) +
  aes(x = variaciones, y = margen, ) +
  geom_line(size = 1.05, colour = "#112446") +
  labs(x = "Incremento propuesto (%)", y = "Margen esperado") +
  theme_minimal() +
  theme(
    axis.title.y = element_text(size = 14L),
    axis.title.x = element_text(size = 14L)
  )
lenx = seq(-3.5,10,0.5)
ggplot(histograma_margen,aes(x = margen_optimo,y = freq_margen) ) +
  geom_point( y = freq_margen, stroke=2,color="#112446",shape=21)+
  geom_segment( aes(x=margen_optimo, xend=margen_optimo, y=0,
yend=freq_margen),color=("#112446"),size=1.5)+
  labs(x = "Margen óptimo", y = "Peso") +
  theme_minimal() +
  theme(axis.title.x = element_text(size = 14L))+
  scale_x_continuous("Margen óptimo", labels = as.character(margen_optimo), breaks = lenx)
 leny = seq(-10,10,0.5)
marg_opt <- scan()
frontera_eficiente = data.frame(variaciones,marg_opt)
```

```
ggplot(frontera_eficiente) +
 aes(x = variaciones, y = marg_opt) +
 geom_line(size = 1.25, colour = "#112446") +
 labs(x = "Variaciones propuestas", y = "Margen esperado") +
 theme_minimal() +
 theme(
   axis.title.y = element_text(size = 14L),
   axis.title.x = element_text(size = 14L)
 )+
 scale_x_continuous("Margen óptimo", labels = as.character(variaciones), breaks = leny)+
 geom_vline(xintercept = 3,colour = "aquamarine",size=2)+
 geom_vline(xintercept = 1,colour = "firebrick",size=2)
```