Anales del Instituto de Actuarios Españoles, 4ª Época, 26, 2020/157-179

ADVANCED ANALYTICS PRICING FOR THE CALCULATION OF POST-COVID19 SCENARIOS IN AUTOMOBILE INSURANCE

TARIFICACIÓN ANALÍTICA AVANZADA PARA CÁLCULO DE ESCENARIOS POST-COVID19 EN LOS SEGUROS DE AUTOMÓVILES

Xenxo Vidal-Llana[*]
Montserrat Guillen[1+]

[*] *PhD(c) de la Universidad de Barcelona, Senior Pricing Actuary de Allianz*
[+] *Catedrática de la Universidad de Barcelona*

Fecha de recepción: 29/09/2020
Fecha de aceptación: 07/11/2020

Abstract

The reduction in mobility during the COVID19 pandemic has led to a reduction in the accident claims rate in motor insurance. Insurance companies will need to calculate pricing scenarios for possible changes in transportation habits, using data from 2020. We show how some Machine Learning methods (decision trees and gradient boosting) can be used to evaluate pricing scenarios and we propose a strategy to correct the circumstances of exposure to risk that have occurred during the pandemic. We conclude that it is possible to use the existing information during the lockdown period provided that changes in the portfolios can be identified and corrected, and assessing whether or not the impact is homogeneous by risk groups.

Key words: predictive modeling, loss, combined ratio, scenarios, Machine Learning.

---

[1] Xenxo Vidal-Llana. Email: xenxovidal@gmail.com. Correspondencia: Montserrat Guillen Email: mguillen@ub.edu. Dirección de ambos autores: Dept. Econometría, Estadística y Economía Aplicada, Riskcenter-IREA, Av. Diagonal, 690, E08034 Barcelona. Telf: +34934037039. Los autores agradecen el apoyo recibido de la Fundación BBVA "Ayudas para equipos de investigación en Big Data", del Ministerio de Ciencia e Innovación PID2019-105986GB-C21 y de ICREA Academia.

DOI: 10.26360/2020_7

Resumen

La reducción de la movilidad durante la pandemia COVID19 ha supuesto una disminución de la siniestralidad en el seguro de automóviles. Las entidades necesitarán realizar escenarios de tarificación para posibles cambios en los hábitos de transporte, usando los datos del ejercicio 2020. Mostramos cómo utilizar métodos de *Machine Learning* (árboles de decisión y *gradient boosting*) para evaluar dichos escenarios y proponemos una estrategia para corregir las circunstancias de exposición al riesgo que se han dado durante la pandemia. Se concluye que es posible utilizar la información existente durante el periodo de confinamiento siempre que se puedan identificar y corregir los cambios en las carteras, viendo si el impacto es homogéneo por grupos de riesgo.

Palabras clave: modelización predictiva, siniestro, ratio combinado, escenarios, aprendizaje automático.

1. Introducción y motivación

La reducción de la movilidad derivada de las medidas de confinamiento durante el año 2020 ha dado lugar a una disminución de la siniestralidad en las carreteras españolas que para los fallecidos en accidente de tráfico ha supuesto una caída del 68% (DGT, 2020). En algunos países como el Reino Unido, se han documentado reducciones medias en el precio del seguro de automóvil a todo riesgo de hasta el 6% en verano de 2020, respecto al máximo observado en el nivel medio de precios alcanzado en 2019 (ver Forbes, 2020). Sin embargo, el retorno a los niveles habituales del uso del transporte privado, o incluso la posibilidad de que se incrementen debido a un posible aumento de la movilidad en vehículos particulares por el temor al contagio en medios de transportes públicos altamente masificados, puede ser un preludio de lo que podría constituir un rebote en la siniestralidad en los meses posteriores a una primera ola de la pandemia.

En este trabajo se presenta una metodología para elaborar escenarios de tarificación que tengan en cuenta (1) que los datos observados durante el periodo de irrupción de la COVID19 no tienen la misma uniformidad en la exposición al riesgo de los vehículos que las de periodos anteriores, debido a los cambios de hábitos derivados de las medidas de restricción provocadas por la pandemia y (2) que existe heterogeneidad geográfica en el impacto de los cambios en la exposición al riesgo de los vehículos asegurados, cuya

complejidad puede no ser posible capturar mediante modelos de tarificación clásicos.

En un trabajo reciente sobre las consecuencias de la pandemia en el sector asegurador, Richter y Wilson (2020) afirman que la crisis de la COVID19 ha puesto de manifiesto la importancia en la redacción de los contratos, y en particular de las cláusulas de exclusión, puesto que son determinantes en pólizas que cubren pérdidas por interrupción de negocios derivadas de causas externas. Aunque no entran en mayor detalle, dichos autores indican también que ahora más que nunca se constata la utilidad del análisis de escenarios para "construir la resiliencia de antemano y planificar acciones de contingencia en escenarios de crisis".

La mayoría de análisis de escenarios de pandemias que se realizaban hasta el día de hoy en las entidades europeas se concentraban en impactos Solo en seguros de vida y salud, teniendo en cuenta los datos de los antecedentes catastróficos históricos, como la epidemia de gripe de 1920, o la peste bubónica del siglo XIV. Ambas situaciones se han considerado actualmente altamente improbables por los avances médicos y de salubridad en las sociedades modernas (Howard, 2020). Lo mismo ha ocurrido con el ejemplo más reciente del SARS que si bien ha servido para fijar un escenario de impacto creíble, no ha tenido la magnitud de la COVID19.

Parece claro que los test de estrés habituales no habían anticipado la realidad que se ha vivido en plena eclosión de la COVID19, ya que básicamente alteraban únicamente las hipótesis de los productos de vida y salud y, a lo sumo, consistían en plantear escenarios que contemplasen un impacto negativo en los mercados financieros derivados del shock pandémico. Según Richter y Wilson (2020) la enorme sorpresa ha sido que la actual amenaza sobre la salud provocada por la COVID19 no Solo ha impactado en las áreas de los seguros de vida y salud, sino que también se ha trasladado a los seguros generales y, principalmente, a través de las pólizas de cobertura de pérdidas de negocio.

Nuestro punto de partida es que la pandemia ha impactado también en los seguros de automóviles, y lo ha hecho debido al enorme repercusión en la movilidad derivada de las restricciones establecidas por los confinamientos. Por esa razón, creemos que la forma de calcular escenarios de tarificación de este seguro obligatorio va a tener una especial relevancia en el periodo post-COVID19.

2. Impacto de la reducción de movilidad en la tarificación.

Una de las aproximaciones a la modelización de la prima de riesgo más comúnmente utilizada es el enfoque de frecuencia y cuantía. En este planteamiento se considera el coste observado durante un periodo temporal separando el número de siniestros y coste de cada uno de ellos. Así pues entra en juego el factor exposición, que marca el porcentaje del año durante el que la póliza ha estado vigente.

Para poder separar el tiempo de exposición y modelizar coste esperado por unidad de exposición, consideramos que la base de cálculo de la prima pura (PR) es el coste medio esperado según el tiempo de exposición que esté una póliza. La prima pura se define del siguiente modo:

$$\mathrm{PR} = \frac{\mathrm{Coste}}{\mathrm{Exposición}}.$$

El enfoque frecuencia y severidad separa el cociente anterior en dos partes, una que se aproximará modelizando el número esperado de siniestros (NStros), y otra que se refiere al coste medio para cada siniestro. Es trivial que es otra definición de la prima pura anterior:

$$\mathrm{PR} = \mathrm{Freq} * \mathrm{Sev} = \frac{\mathrm{NStros}}{\mathrm{Exposición}} \times \frac{\mathrm{Coste}}{\mathrm{NStros}} \qquad (1)$$

Gracias a esta sencilla forma de expresar la prima, se puede derivar que cualquier bajada o subida en la siniestralidad es fácilmente trasladable a precios con solo añadir un multiplicador.

En este trabajo nos centraremos Solo en el modelo de frecuencia y vamos a suponer que el coste medio de un siniestro es constante.

El problema derivado de la situación de pandemia es que el primer factor de la expresión (1) es anormalmente bajo, es decir, no se han producido tantos siniestros como en otros ejercicios dado que la exposición Solo contempla el periodo de vigencia de la póliza y no el uso efectivo del vehículo. Para poder realizar escenarios de tarificación post-COVID19 deberá corregirse este efecto. Por lo tanto, ese primer factor deberá ser multiplicado por una constante superior a 1 a fin de corregir la infraestimación de siniestralidad cuando se creen escenarios de regreso a los niveles normales. Para poder conformar un abanico de escenarios posibles, se debería establecer un

conjunto de factores superiores a 1 que permitiera trasladar las expectativas respecto a la recuperación de los niveles de siniestralidad.

La metodología que se presenta a continuación permite detectar qué parte de la cartera sufre un efecto de la caída de siniestralidad respecto a ejercicios anteriores y cuál es la magnitud de dicha caída. De esta forma, los multiplicadores de corrección para los escenarios de tarificación post-COVID19 no serán uniformes para toda la cartera sino que se adaptarán a los distintos segmentos. La riqueza de los escenarios a contemplar por la tarificación será mucho mayor cuando se tengan en cuenta dichos comportamientos en lugar de efectuar correcciones de ratios de siniestralidad homogéneas para todas las pólizas.

3. Metodología propuesta

3.1 Modelos lineales generalizados

La regresión lineal establece una relación lineal entre la variable respuesta (o variable dependiente) y las variables explicativas (o variables independientes). El modelo de regresión lineal se expresa como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad (2)$$

donde Y_i corresponde a la variable dependiente para al i-ésima póliza (i=1,…,n) y X_{ji}, las correspondientes observaciones de las k variables explicativas, siendo j=1,…,k. El total de observaciones es n y el vector de parámetros a estimar es $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. Se incorpora un término de perturbación ε_i que recoge las desviaciones respecto de la media., y cuya esperanza matemática es cero, de modo que:

$$E(Y_i \mid X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}. \quad (3)$$

Los coeficientes se estiman por mínimos cuadrados ordinarios, que equivalen a un procedimiento de maximización de la verosimilitud cuando se supone que la variable dependiente sigue una distribución normal.

La aplicación de una transformación de la esperanza condicionada de la variable respuesta en la expresión (3) supone una relajación de las hipótesis inicialmente requeridas para este modelo. Además, se obtiene una flexibilidad más amplia a la hora de predecir variables con distribuciones distintas a la gaussiana, por lo que resulta en un nuevo modelo que transformado se denomina modelo lineal generalizado (Kaas et al., 2008). En

el caso que nos ocupa, utilizaremos el modelo de Poisson para modelizar el número esperado de siniestros, que luego podrá utilizarse en (1) para el cálculo de la prima. (Charpentier, 2014 y Denuit, Maréchal, Pitrebois y Walhin, 2007).

3.2 Árboles de decisión

El objetivo principal de los árboles de decisión (Breiman, Friedman, Olshen y Stone, 1984) es dividir la base de datos en subconjuntos y aplicar la media observada de la variable respuesta de cada subconjunto como predicción en aras de minimizar un error objetivo. Debido a su simplicidad, los árboles de decisión pueden ser usados tanto en problemas de regresión (predicción de un número en la recta real), como en problemas de clasificación (predicción de una probabilidad). Cabe notar que tal simplicidad provoca una clara interpretabilidad del modelo.

De esta manera, el árbol se puede interpretar como un conjunto de decisiones tomadas según los valores de los factores de riesgo que derivan en un valor concreto que constituye la predicción.

El hecho de entrenar muchos árboles distintos sobre subconjuntos de los mismos datos y hacer la media de las predicciones se llama *bagging*. El algoritmo más conocido de este estilo es el *Random Forest*, pero no lo utilizaremos en este trabajo.

3.3 Gradient Boosting

Los *Gradient Boosting Machines* (GBM) son modelos propuestos por Friedman (2001) basados en la búsqueda de una función, definida a partir de composición de funciones, que minimice la esperanza de cierto valor de pérdida, o *loss function*.

Se trata nada más y nada menos que de calcular la función gradiente en el conjunto de parámetros, ajustar un modelo de regresión, escoger cómo varían los parámetros para el descenso del gradiente y actualizar la función inicial. En este caso en concreto los modelos ajustados serán árboles de decisión, y lo que hará cada uno de ellos es disminuir el error generado por los anteriores, así, a diferencia de un algoritmo de *bagging*, la predicción realizada por un modelo de *boosting* se deriva de la suma de muchas predicciones de modelos menores. La principal desventaja de los GBM es su baja interpretabilidad, mientras que su mayor ventaja es la capacidad de alcanzar un menor error de predicción.

A partir de la idea que se acaba de exponer, los modelos se van sofisticando con muchas mejoras e implementaciones, como son los *gbm* para R, la implementación de *scikit-learn* para Python, o las más conocidas: *xgboost*, *catboost* y la que usaremos en este trabajo, *LightGBM* (Ke et al., 2017).

Ke et al. (2017) señalan en su artículo que consiguen, mediante una agrupación de sus factores de riesgo, una escalabilidad y eficiencia que hasta ahora no había sido satisfactoria con otros algoritmos de *Machine Learning* cuando el tamaño de los datos es grande. Consiguen así mejorar el tiempo computacional de los algoritmos de *gradient boosting* hasta 20 veces obteniendo casi la misma precisión. Otro factor muy importante en la implementación es la capacidad de tratar variables categóricas, que hasta ahora todos los algoritmos de g*radient boosting* trataban como numéricas.

3.4 Siniestralidad inferior a la habitual en diferentes segmentos

Los algoritmos llamados de *Machine Learning* son útiles en la búsqueda de no linealidades en los comportamientos de los impactos de los factores de riesgo sobre la variable respuesta por su naturaleza.

Para ilustrar los problemas que genera una reducción no homogénea de la siniestralidad, y mostrar la versatilidad de los anteriores modelos, en nuestra ilustración con datos reales supondremos un escenario inicial, donde la siniestralidad es la observada históricamente. Posteriormente la modificaremos aplicando artificialmente una disminución de los siniestros en un segmento de la cartera concreto, y por lo tanto acercándonos a las consecuencias de una disminución de la movilidad que deriva de un confinamiento y que puede afectar Solo a unos asegurados, y más concretamente los que residen en alguna región especialmente afectada por las restricciones de movilidad. Finalmente, plantearemos un tercer escenario con esta misma falta de siniestros en una combinación de factores, ya que Solo se eliminarán siniestros si responden a una combinación no lineal de características, por ejemplo, asegurados de cierta región y a partir de cierta edad del conductor.

El resultado que comparamos en el apartado de análisis de datos es la predicción del número esperado de siniestros en las tres situaciones: siniestralidad habitual, siniestralidad disminuida por un solo factor y siniestralidad disminuida por una combinación de factores. La metodología propuesta tiene que ser capaz de detectar que ha habido una disminución de la siniestralidad y qué factores la causan, para así poder efectuar una

corrección adecuada en los escenarios de tarificación post-COVID19, eliminando el sesgo que se produciría por una corrección de la siniestralidad homogénea, es decir, indiscriminada para todos los asegurados. Esta metodología sería la aplicable a datos recogidos durante la pandemia, a fin de poderlos utilizar en ejercicios posteriores directamente o mediante la configuración de escenarios posibles.

4. Datos y resultados

Los datos utilizados en este trabajo son la tabla freMTPLfreq del paquete R "CASDatasets" (Dutang y Charpentier, 2016). Dichos datos han sido ampliamente estudiados en la predicción de la frecuencia y el coste medio de los siniestros en el mundo académico. Más concretamente la tabla recoge información sobre *Motor Third Party Liabilities*, o Responsabilidad Civil en vehículos a motor, de una aseguradora francesa. Los modelos que se presentan a continuación ajustan con buena precisión los perfiles siniestrales de los conductores para ajustar su prima y poder cubrir el coste esperado de los siniestros que la compañía subsidiará en un futuro. Los datos originales contienen en cada fila una póliza, donde se incluye la variable objetivo, el número de siniestros que se han observado en dicha póliza durante su periodo en vigor, la exposición (porcentaje del año en que la póliza ha estado vigente) y el resto de los factores de riesgo. Distintos ejemplos de trabajos académicos que utilizan los datos aquí presentados para mostrar la metodología de tarificación son los siguientes: Côté et al. (2020) muestran cómo agrupar los datos para eliminar información sensible. Lorentzen y Mayer (2020), Schelldorfer y Wüthrich (2019) y Wüthrich (2019) implementan e interpretan modelos de *Machine Learning* en estos datos. Noll, Salzmann y Wüthrich (2019) además comparan los resultados con modelos clásicos. Su y Bai (2020) estudian la frecuencia y severidad de los siniestros y Počuča, Jevtić, McNicholas y Miljkovic (2020) incluyen propuestas sobre cómo tratar el exceso de ceros (ver también Quazvini, 2019). Existen análisis similares, aunque con otros datos, en los que se tiene en cuenta el problema de la exposición al riesgo de sufrir un siniestro, introduciendo la distancia recorrida por el vehículo asegurado durante un año (Boucher, Côté y Guillen, 2017, Guillen, Ayuso y Pérez-Marin, 2018 y Pesantez-Narvaez, Guillen y Alcañiz, 2019). Otros trabajos incluyen nuevos factores de tarificación que recogen el estilo de conducción (Gao, Wüthrich, y Yang, 2019, Gao, Meng y Wüthrich, 2019).

La muestra utilizada en este trabajo contiene un total de 121,617 pólizas de las 413,169 originales. Se han seleccionado únicamente las que tienen una exposición igual a 1 aunque la misma metodología que va

a emplearse a continuación se puede aplicar cuando existen pólizas no cubiertas durante todo el año. Se ha utilizado un 90% de la muestra como entrenamiento y un 10% como test.

La Tabla 1 muestra la definición de las variables y la Tabla 2, una descripción de los principales estadísticos descriptivos de las variables cuantitativas. Como se observa, los datos reflejan una edad media del conductor de 49.91 años, los vehículos tienen una antigüedad de 8.94 años en promedio y las poblaciones de residencia una densidad media de población de 1.08 miles de habitantes por kilómetro cuadrado. Más de la mitad de los vehículos son de marcas de fabricación francesa (Renault, Nissan o Citroën), la potencia más frecuente es la etiquetada como "*f"* y la mayoría de los vehículos (53.21%) son de gasolina. Finalmente, como se observa en la Tabla 3, la región con más pólizas en vigor es "Centre", con un 52.50% de pólizas.

4.1 Comparación de modelos predictivos

El primer escenario que planteamos es el que muestran los datos originales sin modificación alguna. Los tres modelos que se han estudiado son los GLM (modelos lineales generalizados), los árboles de decisión y *LightGBM* (*gradient boosting*). Comentar que, aunque los tiempos de ejecución son cortos (menos de 5 segundos), el árbol de decisión converge algo más rápidamente que los dos otros modelos.

Tabla 1. Variables de la base de datos

Variable	Descripción
Power	La potencia del vehículo (Ordinal categórica).
CarAge	Antigüedad del vehículo, en años.
DriverAge	Edad del conductor, en años (en Francia se puede conducir a partir de los 18 años).
Brand	La marca del vehículo dividida en los siguientes grupos: A-Renault Nissan y Citroën, B-Volkswagen, Audi, Skoda y Seat, C-Opel, General Motors y Ford, D-Fiat, E-Mercedes Chrysler y BMW, F- Japoneses (excepto Nissan) y coreanos, G-Otros
Gas	Combustible del coche, diésel o gasolina.
Region	Región de la póliza (basada en la clasificación de 1970-2015
Density	Densidad (número de habitantes por km2) de la ciudad que vive el conductor del vehículo
Num_claims	Número de sinestros. Es la variable respuesta

Tabla 2. Estadísticos descriptivos de variables numéricas

Variable	Media	Desv. Est.	Mediana	Mínimo	Máximo
CarAge	8.94	5.46	8	0	100
DriverAge	49.91	14.77	49	18	99
Density	1,077.52	3,436.23	155	2	27,000

Tabla 3. Frecuencias observadas para las regiones de residencia de los conductores

Region	Observaciones	Proporción
Aquitaine	5409	4.4%
Basse-Normandie	3684	3.0%
Bretagne	17502	14.4%
Centre	63798	52.5%
Haute-Normandie	1060	0.9%
Ile-de-France	7647	6.3%
Limousin	1054	0.9%
Nord-Pas-de-Calais	3580	2.9%
Pays-de-la-Loire	11755	9.7%
Poitou-Charentes	6128	5.0%

Se han utilizado todas las variables explicativas en los diferentes modelos, es decir, edad del conductor, antigüedad del vehículo, marca, potencia y tipo de combustible, densidad de la zona de residencia y región. La Tabla 4 muestra el número de siniestros, frecuencias observadas, así como frecuencias predichas por regiones en el primer escenario. En última fila se muestra el RMSE (raíz cuadrada del error cuadrático medio que se calcula en la muestra test) calculado en el 10% de muestra que se utiliza como test.

Tabla 4. Número y frecuencia observada de siniestros y predicción de la frecuencia media para cada región según modelo

	Número	Frecuencia	Frecuencia Predicha		
Region	**Siniestros**	**Observada**	**GLM**	**GLM**	**GLM**
Aquitaine	347	6.42%	6.37%	6.01%	6.09%
Basse-Normandie	214	5.81%	5.86%	5.69%	5.77%
Bretagne	887	5.07%	5.12%	5.38%	5.13%
Centre	2820	4.42%	4.42%	4.46%	4.55%
Haute-Normandie	54	5.09%	5.52%	5.48%	5.59%
Ile-de-France	571	7.47%	7.28%	7.17%	6.85%
Limousin	70	6.64%	6.63%	5.11%	5.42%
Nord-Pas-de-Calais	232	6.48%	6.51%	6.53%	6.12%
Pays-de-la-Loire	660	5.61%	5.60%	5.48%	5.49%
Poitou-Charentes	292	4.77%	4.91%	4.85%	5.00%
RMSE ($\times 10^{-2}$)			22.67	22.69	22.67

En la Figura 1 se presenta el árbol de decisión de éste primer escenario, donde no se ha eliminado ningún siniestro.

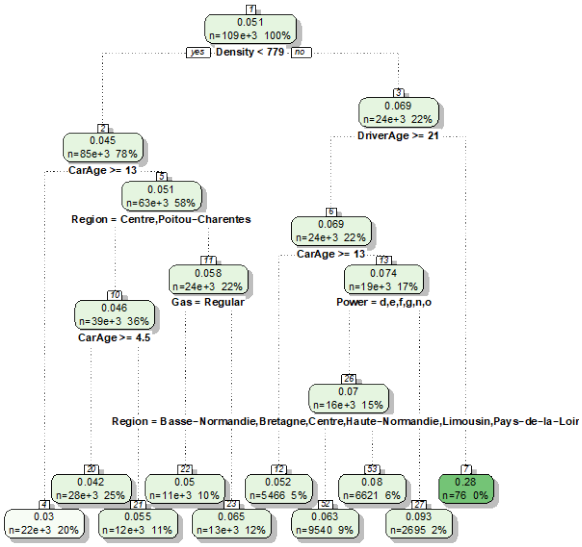

Figura 1: Árbol de decisión. Resultados sobre frecuencia de siniestralidad para los datos originales

La lectura de los números que aparecen en la Figura 1 es simple, en cada nodo terminal (nodos más profundos) del árbol se muestra una casilla con varios valores; arriba se encuentra la predicción de la frecuencia de siniestralidad para la subcartera que pertenece a ese nodo, abajo a la izquierda se presenta el número de pólizas incluidas en el nodo y, a la derecha, el porcentaje respecto al total de la cartera original. En cuanto la interpretación del árbol, se observa que primero separa la cartera en un corte de densidad de la población de residencia del asegurado. Seguidamente realiza una partición ya muy conocida por las aseguradoras (conductores noveles, entre 18 a 20 años, y superiores) y antigüedades más altas del vehículo (para las que se reduce mucho la siniestralidad). Finalmente, el árbol va interaccionando otras variables para mejorar la predicción final.

4.2. Cálculo de escenario: reducción de siniestralidad en una región

El primer escenario alterado que se plantea en nuestro trabajo es una reducción del 70% de los siniestros en la zona de Ile-de-France, para intentar simular los efectos de una reducción de siniestralidad por confinamiento más estricto en la ciudad de París, y dejando intactas las otras zonas. Para efectuar esta reducción se han eliminado aleatoriamente el 70% de los

siniestros observados en la región Ile-de-France. Los resultados de los modelos ajustados se presentan en la Tabla 5 junto a la raíz cuadrada del error cuadrático medio en la muestra de test (RMSE).

Tabla 5. Número y frecuencia observada de siniestros y predicción de la frecuencia media para cada región según modelo y datos con reducción de siniestros en Ile-de-France

	Número	**Frecuencia**	**Frecuencia Predicha**		
Region	**Siniestros**	**Observada**	**GLM**	**Árbol**	**LightGBM**
Aquitaine	347	6.42%	6.44%	5.53%	6.23%
Basse-Normandie	214	5.81%	5.78%	5.62%	5.67%
Bretagne	887	5.07%	5.06%	5.29%	5.07%
Centre	2820	4.42%	4.42%	4.49%	4.45%
Haute-Normandie	54	5.09%	5.06%	5.43%	5.23%
Ile-de-France	146	1.91%	2.03%	2.28%	2.28%
Limousin	70	6.64%	6.49%	5.05%	5.05%
Nord-Pas-de-Calais	232	6.48%	6.50%	5.92%	6.23%
Pays-de-la-Loire	660	5.61%	5.48%	5.37%	5.42%
Poitou-Charentes	292	4.77%	4.75%	4.61%	4.86%
RMSE ($\times 10^{-2}$)			22.21	22.21	22.19

RMSE es la raíz cuadrada del error cuadrático medio que se calcula en la muestra test.

Lo primero que observamos es que todos los modelos han conseguido captar gran parte de disminución en la siniestralidad en Ile-de-France. Si, por tanto, suponemos que este escenario es el año en el que la cuarentena ha sido aplicada, y el primer escenario es el año habitual como el reflejado en la sección anterior, podemos saber por las diferencias de ambas predicciones dónde encontrar las infraestimaciones.

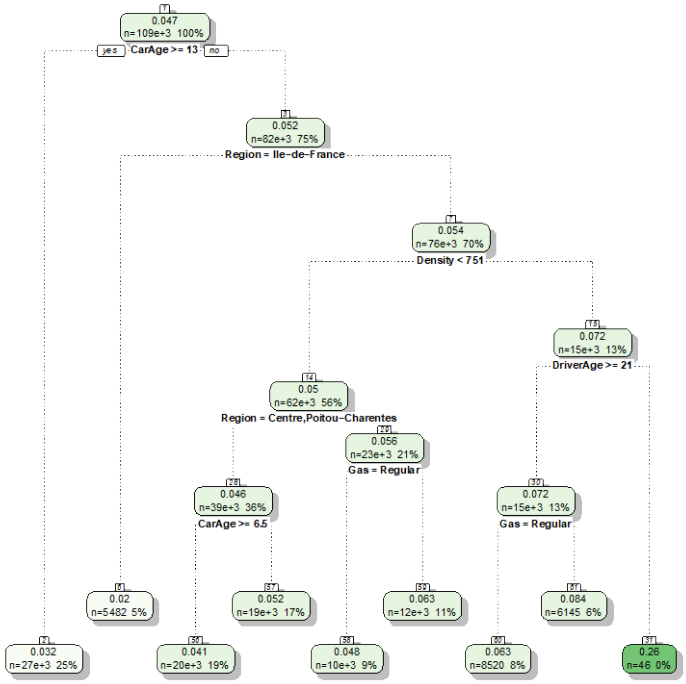

Figura 2: Árbol de decisión. Resultados sobre frecuencia de siniestralidad para los datos con reducción de la siniestralidad en la región Ile-de-France

Es decir, si comparamos la predicción del GLM que hicimos en el apartado anterior para Ile-De-France (Tabla 4) contra la predicción resultante en este nuevo escenario (Tabla 5), vemos una reducción de la frecuencia del 7.28% al 2.03%, por lo que podríamos derivar directamente que aumentar 7.28%/2.03%=3.59, o un 359%, las predicciones para las pólizas de Ile-de-France normalizaría nuestros resultados con los datos estimados a este año simulando un año sin confinamiento. Para el resto de regiones, las predicciones son prácticamente idénticas a las de los modelos aplicados a los datos originales.

El árbol que se muestra en la Figura 2 Solo ha sido modificado en los dos primeros niveles respecto al de la Figura 1, ya que ahora incluye la región y la antigüedad del vehículo como primeros factores diferenciadores, y como sub-árbol se encuentra un árbol muy parecido al de la Figura 1. Notar que no hemos querido aumentar la complejidad del modelo para evidenciar el funcionamiento e intentar comprender qué ocurrirá dentro del LightGBM, de

ahí que surjan diferentes segmentaciones, como la antigüedad del vehículo en primer lugar.

4.3. Cálculo de escenario: reducción de siniestralidad en una región para los conductores de mayores de 45 años

Lo que se aborda a continuación es ver cuáles serían las consecuencias de aplicar disminuciones de la siniestralidad derivadas de la combinación de varios factores. Como ya sabemos, las restricciones por la pandemia no han sido suavizadas en el tiempo ni en el espacio, han sido concretas de zonas y por tiempos limitados, por ejemplo, si se confina un pueblo pero los cercanos no, o si en las primeras fases de la cuarentena Solo pueden salir de la zona perimetral aquellos que vayan a trabajar o que tengan causa justificada. Por tanto, nos preguntamos ¿qué pasaría si la reducción de la siniestralidad Solo la dirigimos, dentro de Ile-de-France, a la asegurados de 45 años o más? Los resultados de entrenar los modelos en una base de datos en la que se ha aplicado una reducción de la siniestralidad del 70% para la región de Ile-de-France y los conductores de 45 o más años se muestran en la Tabla 6.

Como observamos en la Tabla 6, el GLM pierde poder predictivo en lo que respeta a la región Ile-de-France si miramos por grupos de edad. Cuando segmentamos por grupos de edad tiende a sobreestimar a los mayores o iguales de 45 años (predice 4.56%, cuando en realidad se observa 2.2%) e infraestimar a los menores de 45 años (predice 4.81% mientras que se observa 7.53%). Obviamente si se conociera de antemano que hay una reducción de la siniestralidad en la combinación de factores región y edad del conductor, se podría incluir tal interacción en los GLM y ajustaría mejor la siniestralidad, sin embargo, el modelo no es capaz por sí mismo de captar dicha interacción. Este es Solo un mero ejemplo sencillo, ya que en la realidad las interacciones entre factores de riesgo pueden depender de varias variables. Los árboles de decisión sí son capaces de encontrar tales relaciones no lineales entre factores al realizar segmentaciones recursivas.

Tabla 6. Número y frecuencia observada de siniestros y predicción de la frecuencia media para cada región según modelo y datos con reducción de siniestros en Ile-de-France y mayores de 45 años

Region	Grupo de edad	Número Siniestros	Frecuencia Observada	Frecuencia Predicha GLM	GLM	GLM
Aquitaine	≥ 45	191	6.01%	5.97%	5.49%	5.79%
Aquitaine	< 45	156	6.99%	6.81%	5.89%	6.39%
Basse-Normandie	≥ 45	136	5.76%	5.78%	5.60%	5.40%
Basse-Normandie	< 45	78	5.90%	6.47%	6.00%	5.99%
Bretagne	≥ 45	542	4.98%	4.79%	4.87%	4.88%
Bretagne	< 45	345	5.22%	5.46%	5.05%	5.39%
Centre	≥ 45	1732	4.30%	4.23%	4.42%	4.36%
Centre	< 45	1088	4.63%	4.76%	4.48%	4.67%
Haute-Normandie	≥ 45	30	5.04%	4.84%	5.43%	4.79%
Haute-Normandie	< 45	24	5.16%	5.30%	5.74%	5.74%
Ile-de-France	**≥ 45**	**93**	**2.20%**	**4.56%**	**5.71%**	**3.75%**
Ile-de-France	**< 45**	**258**	**7.53%**	**4.81%**	**6.02%**	**6.44%**
Limousin	≥ 45	42	7.08%	6.65%	5.14%	5.16%
Limousin	< 45	28	6.07%	7.39%	5.35%	5.33%
Nord-Pas-de-Calais	≥ 45	118	6.42%	5.87%	5.69%	5.76%
Nord-Pas-de-Calais	< 45	114	6.55%	6.67%	6.11%	6.35%
Pays-de-la-Loire	≥ 45	349	5.42%	5.37%	5.42%	5.27%
Pays-de-la-Loire	< 45	311	5.85%	5.94%	5.71%	5.79%
Poitou-Charentes	≥ 45	170	4.76%	4.40%	4.55%	4.65%
Poitou-Charentes	< 45	122	4.77%	4.95%	4.63%	4.95%
RMSE ($\times 10^{-2}$)				22.43	22.46	22.41

RMSE es la raíz cuadrada del error cuadrático medio que se calcula en la muestra test

Como podemos observar en la Figura 3, el árbol de decisión es capaz de captar la reducción de la siniestralidad de esta situación. Gracias a la intrerpretabilidad que los caracteriza, podemos ver cómo se modifica el árbol según cada escenario para intentar entender qué ocurre dentro del *gradient boosting*.

En el último escenario estudiado, los siniestros de mayores de 45 de Ile-de-France se han visto reducidos, la condición queda sumergida al fondo del árbol como puede observarse en la Figura 3, y por ello, y a diferencia del árbol representado en la Figura 2, se observa como aparece *a posteriori* la separación por la edad del conductor. Notar que la separación por edad se produce a los 25 años ya que el árbol considera durante su entreno que es mejor corte que 45 años para reducir el error generado. Según se muestra en la Tabla 6, se predeciría en media una siniestralidad del 5.71% y 6.02%

respectivamente para los mayores o iguales de 45 años y el resto de pólizas de Ile-de-France, lo que va acorde con la disminución de siniestralidad de esta región, que depende del factor edad. Aunque la tendencia ha sido captada ligeramente, un solo árbol no consigue la precisión y segmentación de un algoritmo de *boosting*, ya que las diferencias de predicciones que se obtienen dentro de la región Ile-de-France para el *LightGBM* evidencian su complejidad y precisión, 3.75% para mayores o iguales a 45 años y 6.44% para los menores de 45 años.

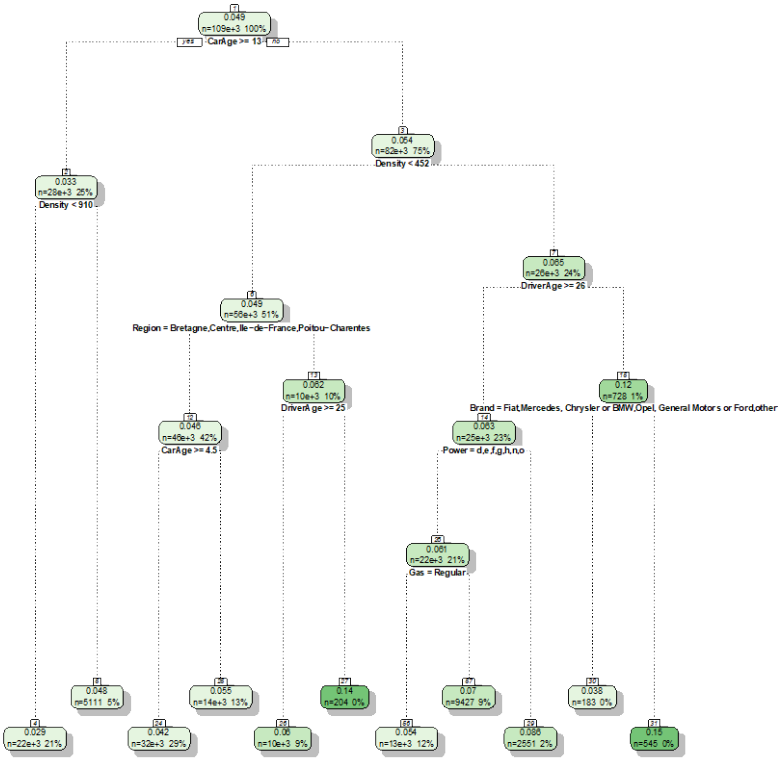

Figura 3: Árbol de decisión. Resultados sobre frecuencia de siniestralidad para los datos con reducción de la siniestralidad en la región Ile-de-France para conductores con edad igual o superior a 45 años

4.4 Implicaciones para la tarificación y escenarios post-COVID19

Los ejemplos mostrados en las secciones anteriores nos proporcionan una herramienta para analizar los escenarios de tarificación en situaciones post-COVID19. En el caso de una cartera real, el apartado 4.1 correspondería a

los resultados de un año sin cambios en la exposición al riesgo, es decir, sin reducciones de la siniestralidad. Este resultado podría obtenerse analizando datos hasta marzo de 2020 para el caso de entidades españolas. En cambio, los resultados del apartado 4.3 indican que, al utilizar datos del ejercicio 2020 completo en los que se habrían producido disminuciones sustanciales de la siniestralidad en algún segmento (aquí se ha simulado Ile-de-France y conductores mayores o iguales a 45 años), los modelos de *Machine Learning* son capaces de detectar la singularidad de la disminución de la siniestralidad. La Tabla 6 muestra que existen notables discrepancias entre las predicciones de siniestralidad para Ile-de-France y los diferentes grupos de edad si se usan modelos predictivos clásicos (GLMs) y modelos basados en árboles.

La consecuencia de los anteriores resultados es que una entidad aseguradora que utilizara los modelos más avanzados, sería capaz de detectar dónde y para qué segmento concreto se ha producido un cambio en la siniestralidad debida a las restricciones de movilidad de COVID19. La forma de detectar dichos comportamientos anormales sería a través de la comparación de árboles de decisión resultantes de un ejercicio normal (apartado 4.1) y un ejercicio alterado (apartado 4.3), por lo que pasan a ser imprescindibles para conocer en qué ramas se producen las mayores diferencias de siniestralidad. Una vez identificadas las interacciones, se debería decidir si aplicar en dichos segmentos la frecuencia siniestral observada en 2020 y por lo tanto más reducida, o, por el contrario, realizar algún escenario de vuelta a la normalidad con frecuencias parecidas a las previas a la situación de pandemia.

La Tabla 7 muestra el cociente entre la predicción de siniestralidad del escenario con movilidad reducida inicial frente al inicial, analizando las diferentes regiones y los dos grupos de edad (menores de 45 frente al resto). Un resultado negativo indica reducción de la siniestralidad. Cuanto mayor sea la magnitud de dicho cociente, mayor es el cambio que debería introducirse en la tarificación para reflejar el comportamiento observado. En el caso de querer realizar escenarios de vuelta a la normalidad, sería en los segmentos con un cociente más elevado en valor absoluto donde resultaría necesario efectuar hipótesis más intensas de cambios de comportamiento. En cambio, en otras zonas y grupos de la cartera, si no hay un cociente excesivamente elevado, sería perjudicial realizar un cambio de escenario porque éste podría haberse debido a un cambio en el mix de negocio.

Tabla 7. Reducción porcentual* en la predicción de la frecuencia siniestral para cada región y grupo de edad según el modelo utilizado, al comparar el escenario original y el que resulta de los datos con reducción de siniestros en Ile-de-France y mayores de 45 años

Region	Grupo de edad	GLM	Árbol	LightGBM
Aquitaine	≥ 45	-3.62%	-7.11%	-2.76%
Aquitaine	< 45	2.64%	-4.17%	1.69%
Basse-Normandie	≥ 45	0.87%	-0.85%	-5.37%
Basse-Normandie	< 45	6.44%	3.79%	1.64%
Bretagne	≥ 45	-3.76%	-8.21%	-2.36%
Bretagne	< 45	1.65%	-8.00%	0.78%
Centre	≥ 45	-2.04%	-0.89%	-2.45%
Centre	< 45	3.83%	0.38%	-0.29%
Haute-Normandie	≥ 45	-10.52%	0.79%	-11.65%
Haute-Normandie	< 45	-6.31%	2.66%	-1.07%
Ile-de-France	**≥ 45**	**-37.02%**	**-20.16%**	**-43.60%**
Ile-de-France	**< 45**	**-34.43%**	**-16.25%**	**-9.18%**
Limousin	≥ 45	3.02%	1.61%	-3.88%
Limousin	< 45	7.80%	3.45%	-2.97%
Nord-Pas-de-Calais	≥ 45	-6.52%	-12.41%	-3.12%
Nord-Pas-de-Calais	< 45	-1.27%	-6.90%	0.89%
Pays-de-la-Loire	≥ 45	-1.79%	-0.09%	-0.79%
Pays-de-la-Loire	< 45	3.05%	2.93%	1.43%
Poitou-Charentes	≥ 45	-8.04%	-5.12%	-5.81%
Poitou-Charentes	< 45	-2.66%	-5.72%	-2.75%

* La reducción porcentual se calcula como el cociente entre la predicción obtenida en los datos con reducción de la siniestralidad (Tabla 6), dividida por la predicción obtenida en los datos originales (Tabla 4) menos 1. En negrita se destaca la región Ile-de-France.

Si considerásemos el caso más sencillo, donde corregimos la siniestralidad observada a toda nuestra cartera por igual, tendríamos que inicialmente observamos un 5.05% de siniestralidad media, y tras la reducción forzada en Ile-de-France para los asegurados de 45 o más años, esta se convierte en un 4.87%.

Una manera posible de compensar esta falta de siniestros a la espera de un retorno a la normalidad cuando se están realizando escenarios posibles, sería modificar el precio suponiendo que la frecuencia siniestral ha sido anormalmente baja, 4.87% y aumentarla artificialmente en un escenario pesimista (mayor siniestralidad) a fin de aproximar lo máximo posible el comportamiento visto el año previo a la pandemia. En otras palabras, este tratamiento indicaría que si se observa una siniestralidad en 2020 que es de 4.87%, mientras que normalmente la siniestralidad venía siendo de 5.05%, se optaría por un escenario de incremento de precios en 2021, derivado de

dicha diferencia e igual a 5.05%/4.87%=1.037, para corregir lo que sería el reflejo fiel de lo ocurrido en 2019.

Esta forma de plantear el escenario post-COVID19, no sería justo para esta cartera, ya que estaríamos sobre-tarificando muchas pólizas, para las que no es cierto que su comportamiento en 2020 haya sido de tan baja siniestralidad. En cambio, no corregiríamos lo suficiente y estaríamos infra-tarificando las pólizas de Ile-de-France para mayores de 45 años. Es en la Tabla 7 donde se puede observar cómo podríamos detectar las combinaciones que más se han visto afectadas por la pandemia, y también corroboramos los resultados obtenidos en la Tabla 6, donde el uso de modelos clásicos produciría una segmentación en cuanto a región se refiere, pero no captaría que los menores de 45 no se han visto tan afectados por esta reducción de siniestralidad.

5. Conclusiones

Como es bien sabido, en el entorno asegurador han sido usados históricamente los modelos lineales generalizados, que han supuesto una forma de predecir la frecuencia de siniestralidad con una precisión más que notable a la par que interpretabilidad de los resultados. El hecho de poder comprender a partir de los coeficientes estimados de un GLM cuánto aumenta el precio y a causa de qué factor de riesgo se producen cambios en la prima de riesgo supone una ventaja para comprender los mecanismos que determinan el comportamiento siniestral de una cartera. Los modelos predictivos basados en GLM facilitan la difusión del mensaje sobre cuáles son los factores de riesgo y cuál es su impacto a todos los rangos de una compañía. Sin embargo, nos encontramos en una época donde la recopilación de datos masivos es nuestro día a día, y donde cada día más observamos que la competitividad entre empresas es un factor clave del éxito, es por eso que los algoritmos de *Machine Learning* aportan una renovación de los métodos clásicos de tarificación.

La pregunta "¿Se prefiere interpretabilidad o precisión?" ha de ser tomada con muchas comillas. Como se ha mostrado en los resultados, las interacciones de los factores de riesgo son fácilmente captadas por modelos de ML, mientras que los modelos clásicos requieren de un análisis descriptivo más detallado en aras de disminuir estos errores, y la introducción de los efectos cruzados no es automática. Si los efectos se conforman a través de varios factores, se complica la interpretabilidad del efecto neto de cada covariable. Los métodos de *Machine Learning* son más flexibles en el sentido de poder captar elementos de mayor complejidad en la

interacción entre factores, aunque como contrapartida pueden no ser fácilmente interpretables.

La regulación exige actualmente poder explicar los algoritmos usados para tarificar un seguro en el mercado español, aunque podemos diferenciar modelos usados para el entorno comercial y modelos usados en el entorno técnico, de forma que pueden inducir elementos de control de riesgos que hagan modificar criterios comerciales según resultados derivados de modelos internos.

Otro punto a mencionar es la tendencia de algunos algoritmos de ML a dar predicciones extremas. Mientras que los GLM modifican la media de los datos y finalmente crean una distribución poco sensible a cambios, los algoritmos de ML suelen extremar las predicciones. Esto se observa en los árboles de decisión de la sección anterior. No hay suavidad en las predicciones, se corre el peligro de al madurar una póliza un año y, al cambiar edades, se cambia drásticamente el precio asignado. Esto queda solucionado en los algoritmos de *gradient boosting*, donde la sucesión masiva de árboles provoca una suavización en las predicciones bajo cambios menores en los factores de riesgo.

El punto que queremos mostrar en este estudio es que bajo cambios no lineales en una cartera, como puede ser el confinamiento de asegurados de cierta edad en ciertas ciudades o restricciones horarias, los algoritmos de ML pueden ayudar a comparar un año con los anteriores, permitiendo así tarificar de una manera más precisa a los clientes, por lo que seríamos capaces de, usando datos del año en curso, predecir la siniestralidad futura y tarificar las pólizas para el próximo año habiendo simulado la siniestralidad esperada en éste si no hubiese ocurrido la pandemia de la COVID-19, al mismo momento que habríamos mantenido los distintos perfiles siniestrales de la cartera bien tarificados.

6. Referencias

Boucher, J-P., Côté, S. y Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models, *Risks*, 5(4), 54. https://doi.org/10.3390/risks5040054

Breiman L., Friedman J. H., Olshen R. A. y Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth. Belmont, California.

Charpentier, A. (2014). *Computational Actuarial Science with R*. CRC press, Boca Raton.

Côté, M. P., Hartman, B., Mercier, O., Meyers, J., Cummings, J. y Harmon, E. (2020). *Synthesizing property & casualty ratemaking datasets using generative adversarial networks*. ArXiv preprint arXiv:2008.06110.[Fecha del último acceso: 25.09.2020] https://arxiv.org/pdf/2008.06110.pdf

Denuit, M., Maréchal, X., Pitrebois, S. y Walhin, J.F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons, Berlin.

Dirección General de Tráfico (2020). *Accidentalidad en Semana Santa. Se reducen un 52% los fallecidos*. [Fecha del último acceso: 25.09.2020] http://revista.dgt.es/es/noticias/nacional/2020/04ABRIL/0415siniestralidad-vial-en-Semana-Santa.shtml#.X1e3uHnHzIU

Dutang, C. y Charpentier, A. (2016). *CASdatasets. R package version, 1-0*. [Fecha del último acceso: 25.09.2020] http://cas.uqam.ca/pub/R/web/CASdatasets-manual.pdf

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Gao, G., Meng, S. y Wüthrich, M.V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2019(2), 143-162.

Gao, G., Wüthrich, M.V. y Yang, H. (2019). Evaluation of driving risk at different speeds. *Insurance: Mathematics and Economics*, 88, 108-119.

Guillen, M., Nielsen, J.P., Ayuso, M. y Pérez-Marin, A.M. (2018). The use of telematics devices to improve automobile insurance rates. *Risk Analysis* 39(3), 662-672.

Howard, J. (2020). Plague was one of history's deadliest diseases—then we found a cure, *National Geographic*, [Fecha del último acceso: 25.09.2020] https://www.nationalgeographic.com/science/health-and-human-body/ human -diseases/the-plague/.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. y Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154). Editado por I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan y R. Garnett. Proceedings of "Neural Information Processing Systems 2017." Long Beach, CA, 4-9 Diciembre, 2017. [Fecha del último acceso: 25.09.2020] http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision

Lorentzen, C. y Mayer, M. (2020). *Peeking into the black box: an actuarial case study for interpretable machine learning*. SSRN. http://dx.doi.org/10.2139/ssrn.3595944

Noll, A., Salzmann, R. y Wüthrich, M. V. (2020). *Case study: French motor third-party liability claims*. SSRN 3164764. http://dx.doi.org/10.2139/ssrn.3164764

Pesantez-Narvaez, J., Guillen, M. y Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, 7(2), 70. https://doi.org/10.3390/risks7020070

Počuča, N., Jevtić, P., McNicholas, P.D. y Miljkovic, T. (2020). Modeling frequency and severity of claims with the zero-inflated generalized cluster-weighted models. *Insurance: Mathematics and Economics*, 94, 79-93.

Forbes (2020). *Car Insurance Hike Looms Post-Covid, Maverick Demands Pricing Reform* [Fecha del último acceso: 25.09.2020] https://www.forbes.com/sites/advisoruk/2020/07/07/car-insurance-hike-looms-post-covid-maverick-demands-pricing-reform/#53709af369cb

Richter, A. y Wilson, T.C. (2020). Covid-19: implications for insurer risk management and the insurability of pandemic risk. *The Geneva Risk and Insurance Review*. En prensa. https://doi.org/10.1057/s10713-020-00054-z

Schelldorfer, J. y Wüthrich, M.V. (2019). *Nesting classical actuarial models into neural networks*. Available at SSRN 3320525. http://dx.doi.org/10.2139/ssrn.3320525

Su, X. y Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PloS one*, 15(8), e0238000. https://doi.org/10.1371/journal.pone.0238000

Wüthrich, M.V. (2019). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal*, 1-24.