MACHINE LEARNING ALGORITHMS FOR AUTO INSURANCE FRAUD DETECTION

ALGORITMOS DE MACHINE LEARNING PARA LA DETECCIÓN DEL FRAUDE EN EL SEGURO DE AUTOMÓVILES

Badal Valero, Elena[1], Sanjuán Díaz, Andrés y Segura Gisbert, Jorge

Facultat d´Economia, Universitat de València. Avenida de los naranjos SN. 46022 (Valencia).

Fecha de recepción: 18/06/2020
Fecha de aceptación: 02/11/2020

Abstract

Automobile insurance fraud has increased considerably in recent years, undoubtedly boosted by the economic crisis. This significant increase in fraudulent files and the new requirements of the regulations associated with Solvency II lead to greater control and allocation of resources against fraud by insurers. For these reasons, the importance of the use of prediction techniques for the detection of suspicious accidents is more than justified.

In this paper, we present several methodologies with statistical foundation and automatic learning algorithms that enable the analysis and detection of such claims.

Keywords: Fraud, Machine Learning, Insurance car, Risk

Resumen

El fraude en el seguro de automóvil ha aumentado considerablemente en los últimos años, indudablemente impulsado por la crisis económica. Este

[1] Buzón de correspondencia Elena.badal@uv.es, colaboradora de la Cátedra Deblanc, convenio marco entre la Consellería de Justicia, Interior i Administracions Públiques de la Generalitat Valenciana y la Universitat de València, y de la Consellería d'Innovació, Universitats, Ciència i Societat Digital de la Generalitat Valenciana en el marco del proyecto AICO/2019/05.

DOI: 10.26360/2020_2

incremento significativo del número de reclamaciones fraudulentas, así como los nuevos requerimientos asociados con Solvencia II, conducen a un mayor control y asignación de recursos contra el fraude por parte de las aseguradoras. Por estas razones, la importancia del uso de avanzadas técnicas de predicción para la detección de accidentes sospechosos está más que justificada.

En este artículo, se presentan diversas metodologías de base estadística y algoritmos de aprendizaje automático que permiten el análisis y la detección de tales afirmaciones.

Palabras clave: Fraude, Aprendizaje Automático, Seguro de Automóvil, Riesgo.

1. Introducción

El fraude en el seguro de automóvil es un problema inherente del sector asegurador, con una clara correlación al *momentum* del ciclo económico y especial incidencia en zonas con un desempleo basal. El aumento de expedientes fraudulentos, las nuevas exigencias regulatorias asociadas a Solvencia II y una creciente conciencia de las entidades, ha conllevado a un mayor control y asignación de recursos para la lucha contra el fraude. En este sentido, la transformación que estamos presenciando en la actualidad bajo el paradigma disruptivo de las tecnologías de la información, está coadyuvando la utilización de técnicas, recursos estadísticos y algoritmos de aprendizaje automático para la detección y prevención del fraude en el seguro del automóvil.

El reconocimiento de esta problemática implica la adopción de sistemas de prevención y la identificación de patrones de comportamiento de distintos aspectos que conciernen al asegurado en su conducta frente al siniestro. Así, se define el fraude como *"la situación que se produce cuando el propio asegurado o beneficiario ha procurado intencionadamente la ocurrencia del siniestro o exagerado sus consecuencias con ánimo de conseguir un enriquecimiento injusto a través de la indemnización que espera lograr del asegurador"* (ASEPEG, 2020). Por tanto, son actividades ilegales en las que el asegurado trata de conseguir un lucro o beneficio el cual no le pertenece por contrato. De forma genérica, la conducta se encapsula en una exagerada reclamación de daños (materiales y/o corporales), inexistencia del siniestro o simplemente tramas organizadas que coexisten con distintos roles en aras de

cometer el hecho delictivo. En algunas ocasiones, las reclamaciones son de escasa cuantía lo que dificulta su detección debido a la gran cantidad de siniestros que son tramitados diariamente.

Diversos investigadores, como Cummings y Tennyson (1996), Picard (2000) y Crocker y Tennyson (2002) han analizado el comportamiento deshonesto de los asegurados y su posible cuantificación. Del mismo modo, en materia de técnicas de detección del fraude, se han sucedido la publicación de trabajos que implementan algoritmos de aprendizaje automático para su modelización y predicción. En particular, la problemática del fraude ha sido tratada con modelos de regresión lineal múltiple, redes neuronales artificiales, modelos de elección discreta y la teoría de los conjuntos borrosos.

Las técnicas de regresión se han empleado para estudiar el fraude desde un prisma yuxtapuesto, utilizando variables independientes de índole económica, según plantean Cummins y Tennyson (1996). Belhadji y Dionne (1997) proponen un modelo probit simple con el que pretendían estimar la probabilidad de fraude *versus* no fraude, detectado o sospechado. Los resultados conseguidos son cotejados con un modelo de regresión lineal y la variable objetivo es la cota de sospecha del expediente en una escala comprendida de 0-10 puntos, buscando una articulación automática de detección para alcanzar la eficiencia en la investigación de siniestros sospechosos bajo una dupla coste versus beneficio obtenido. Por su parte, Artís, Ayuso y Guillén (1999) presentan modelos de elección discreta mediante la incorporación de la influencia del asegurado y las características de la reclamación, correspondientes a una muestra de expedientes procedentes del mercado asegurador español.

Para el estudio de las distintas tipologías siniestros fraudulentos, se proponen modelos logísticos multinomiales, basados en la idea de maximizar la función de utilidad esperada, determinando qué variables debe priorizar la entidad aseguradora para la tramitación del expediente (Ayuso & Guillén, 1999).

La utilización de la teoría de los conjuntos borrosos para la detección del fraude fue presentada por Derrig y Ostaszewski (1995**),** realizando una concentración de siniestros en función de la sospecha de fraude en una escala de 0-10. La valoración realizada para los diferentes elementos del siniestro permitirá clasificar el expediente. Por su parte, las redes neuronales se han empleado como sistemas de detección del fraude bajo las hipótesis subyacentes de establecer qué modelos de siniestros similares presentarán

niveles de sospecha análogos, asumiendo que cada uno de los indicadores considerados tendrá igual importancia a la hora de explicar la existencia de fraude, Brockett, Xia y Derrig (1995).

2. El fraude en el seguro del automóvil

En los seguros de automóviles, todo poseedor de un vehículo a motor está obligado a contratar y mantener en vigor una póliza de seguro que cubra la Responsabilidad Civil del conductor que se derive de los daños, tanto personales como materiales, ocasionados a terceras personas como consecuencia de un hecho de la circulación[2].

Los asegurados que incurren en fraude pretenden un objetivo episódico: pasar desapercibidos, no ser descubiertos y obtener cierto lucro individual y particular. Se realizan toda clase de ocultaciones y complicidades sigilosas entre los intervinientes provocando que la detección del fraude no sea nada fácil de descubrir.

El sector asegurador ha incrementado notoriamente sus esfuerzos de detección del fraude en los últimos años. El fraude se presenta en todos los ramos y líneas de negocio, aunque en cada uno se desarrolla de una manera distinta. La escasez de información del asegurador (información asimétrica) sobre la conducta de su cartera de clientes es uno de los principales problemas que presenta el sector, y en especial el ramo de automóviles. Los principales tipos de fraude, según datos de ICEA (2018), son los siguientes:

- **Ocultación de daño o preexistencia del daño**. Hace referencia al daño o lesión producida con anterioridad a la formalización del contrato con información asimétrica por parte del asegurado.
- **Desproporción en la reclamación**. Suele producirse en siniestros con presencia de daños corporales, exagerando dolencias para solicitar una mayor indemnización en función de los daños sufridos.
- **Simulación de un siniestro de forma deliberada.** El siniestro realmente no ha ocurrido o se realiza premeditadamente con la

[2] Real Decreto Legislativo 8/2004, de 29 de octubre, por el que se aprueba el texto refundido de la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor.

intención de conseguir un lucro a costa de la entidad aseguradora.

- **Exclusión de cobertura**. El asegurado pretende recibir una compensación cuando el daño no está cubierto por la póliza de seguro.
- **Falta de nexo causal.** No hay relación de causalidad entre el daño originado y el siniestro finalmente declarado.
- **Fraude en la suscripción del riesgo**. Está vinculado con el uso de datos adulterados, referidos tanto al propio asegurado (edad, antigüedad del carnet,…) como los apuntados al vehículo asegurado.
- **Falsedad de facturas, minutas u otro tipo de documentación.** Falsificación de documentos con el objetivo de acrecentar el coste de los daños producidos por el expediente.

En cualquier caso, el 63,6% de los expedientes de fraude detectados se concentra en el seguro del automóvil, seguido con un 29,4% por los ramos masa de diversos y la Responsabilidad Civil General (ICEA, 2018). Particularmente, para la detección y prevención del fraude en el seguro del automóvil deberemos tener en consideración los aspectos que se detallan a continuación:

- **El porcentaje de siniestros fraudulentos es limitado**. Bajo un criterio estadístico-actuarial, el número de siniestros etiquetados como fraude es ínfimo frente al número de siniestros legítimos. Esta cuestión es identificada como un problema de datos desequilibrados donde concentraremos los esfuerzos de identificación de las acciones minoritarias de determinadas instancias.
- **Correcta clasificación de los siniestros fraudulentos**. De la calidad de los datos dependerá el resultado final de los algoritmos de aprendizaje automático.
- **El fraude en el seguro del automóvil presenta gran aclimatación a las particularidades del entorno.** La existencia de verdaderas tramas con múltiples nexos y beneficios correlativos: Talleres, abogados especializados, clínicas de rehabilitación,… De este modo, los sistemas de monitorización y prevención del fraude deberán ser enmendados para una utilidad efectiva.
- **Las metodologías estadísticas deben favorecer la toma de decisiones.** Debe primar la cultura de análisis y la toma de

decisiones fundamentada en los datos simplificando el trabajo del tramitador de siniestros.

El proceso y la transformación en la prevención del fraude implican establecer cuatro actividades principales, según Van Vlasselaer, Eliassi-Rad, Akoglu, Snoeck y Baesesns (2017).

1. **Detección y prevención del fraude**. Aplicación de las distintas metodologías en siniestros nuevos y estipular un riesgo de siniestro fraudulento para cada uno de ellos.
2. **Proceso de investigación del fraude**. No se puede sustituir el juicio experto del tramitador de siniestros dada la elevada complejidad del proceso de fraude, aunque en muchas ocasiones resulte muy difícil de demostrar. Este proceso consume cuantiosos recursos (no es inmediato) y hace que la tramitación se dilate en el tiempo, resultando en ocasiones muy compleja su demostración fehaciente. El desarrollo de acciones formativas a los tramitadores es fundamental a medio y largo plazo.
3. **Siniestro fraudulento confirmado**. El siniestro es fraudulento de forma parcial o total. Los equipos de verificación y lucha contra el fraude determinarán el grado de fiabilidad del modelo o si necesitamos una modificación del mismo.
4. **Elementos de prevención en el futuro**. Qué mecanismos de prevención del fraude son útiles y, dentro de todo el universo de medidas, determinar el grado de eficiencia.

El fraude es un fenómeno activo y dinámico, lo que obliga a los modelos de prevención y detección del fraude a estar en continua evolución. Todo ello influirá en el proceso de detección de manera decisiva.

3. Descripción de la Base de Datos

Se posibilita, para la realización del estudio, una base de datos de pólizas de automóviles de una Entidad Aseguradora con presencia en el ramo de automóviles dentro de su portfolio comercial. La base de datos es un fichero de texto plano (.csv) compuesta por determinadas variables. Este formato permite trabajar con los datos en diferentes programas de análisis estadístico sin necesidad de realizar modificaciones. En particular, se pretende que los datos estén en los formatos apropiados, especificaciones correctas y variables bien etiquetadas. Para la elaboración de la base de datos se ha decidido extraer una muestra del total de siniestros ocurridos en un período concreto de tiempo de 60.414 expedientes.

Es conveniente resaltar que el desarrollo de un estudio completo del fraude esboza la dificultad de establecer muestras de siniestros fraudulentos lo suficientemente significativas como para poder inferir el comportamiento del conjunto de la población (Ayuso et al., 1999).

La base de datos recopilada comprende información relativa a las particularidades inherentes de los productos en vigor, de los distintos canales de distribución, características propias de los asegurados, variables sociodemográficas e información referente a siniestros acaecidos en el pasado. Se realiza una perspectiva general de los valores únicos para cada variable, datos ausentes, duplicados o erróneos, así como la presencia de ceros en términos de frecuencia absoluta de cada variable. Esta etapa de depuración, es de vital trascendencia para los análisis a realizar a posteriori. No se puede aguardar que un modelo sea aceptable si los datos de génesis o partida no son buenos (Silver, 2014). De forma paralela, se realiza una modificación o generación de nuevas variables a partir de determinada información existente, una transformación del formato fechas para su tratamiento y la consideración de variables cualitativas en factores.

El aspecto esencial del análisis es entender qué se considera fraude de un siniestro en el establecimiento de la variable dependiente. De este modo, se procedió a clasificar en una misma categoría a los siniestros fraudulentos confirmados y siniestros con profundas sospechas de presunción de fraude. Por otro lado, se encuentran los siniestros clasificados como normales o legítimos.

Los siniestros fraudulentos representan un conjunto pequeño del total de la actividad aseguradora. Así, si dichos siniestros son el 5% de las acciones fraudulentas, un modelo que no tuviera en cuenta estas circunstancias tendría un porcentaje de acierto del 95%. Resulta evidente que en presencia de datos tan desequilibrados, la clase minoritaria es la más importante (Yen & Lee, 2009).

Existe un sinfín de argumentos que inciden en el análisis de los datos desequilibrados y su impacto en la precisión de los algoritmos de aprendizaje automático. Como exponen López, Fernández, García, Palade, y Herrera (2013) se distinguen los siguientes motivos:

- La distribución dispar de la variable objetivo. La presencia de siniestros sin indicios de fraude cuenta con una elevada presencia de

registros, formando un grupo más uniforme y con menor variabilidad que el siniestro con fraude positivo.

- Los algoritmos tienen como uno de sus principales objetivos disminuir el error general al que la clase minoritaria aporta muy poco y donde existe una tendencia de la clasificación hacia la clase mayoritaria. Hemos de tener en cuenta que estas metodologías y algoritmos parten del supuesto de clases equilibradas.
- Los errores tienen la misma ponderación para las distintas clases y no tiene porqué ocurrir. De este modo, los errores tienen que ser valorados en función de su relevancia.

Para el tratamiento de la problemática anterior, se conocen diversas técnicas que favorecen el balanceo de los datos, como las propuestas por Chawla, Bowyer, Hall y Kegelmeyer (2002) o Lunardon, Menardi y Torelli (2014). Además, existe cuantiosa evidencia científica que demuestra que los resultados obtenidos con una base de datos balanceada mejora los distintos clasificadores en contraposición con la presencia de datos desequilibrados, dado que se evita que los modelos proporcionen predicciones sesgadas hacia la clase mayoritaria (Chawla, 2005; He y García, 2009). En este trabajo, se realiza un muestreo de los datos mediante el balanceo de las clases a través de la adición o eliminación de registros, según exponen He y García (2009) o Guo et al. (2017).

La base de datos creada ad hoc mediante muestreo se compone de 1.048 siniestros, de los que 588 siniestros tendrían la condición de legítimos (56%) y 460 responden a siniestros fraudulentos (44%) o con presunción de fraude, lo que evita la presencia de datos desequilibrados en la base de datos empleada.

Las variables más importantes que se generaron fueron las siguientes:

- **Canal de contratación:** Variable que proporciona información sobre quién ha formalizado el contrato (Corredor, Agente exclusivo, Agente Vinculado,…), mediante una codificación interna por parte de la entidad aseguradora. Para facilitar su tratamiento y análisis se realiza una transformación de la variable en dicotómica. De este modo, tendremos valor 1 si la póliza ha sido formalizada por un corredor y valor 0 en caso contrario.
- **Antigüedad de la póliza:** Se indica la antigüedad del contrato en el momento de acaecer el siniestro representado en valores absolutos.
- **Variaciones del contrato:** En algunas ocasiones la póliza incorpora variaciones desde el mismo momento de su emisión. Así, las variaciones hacen alusión a cambios en la forma de pago,

incorporación o eliminación de garantías, cambios en el domicilio o lugares de residencia, etc. La inexistencia de cambios en el contrato supone tomar como valor 0.

- **Forma de pago:** Establece como se efectúa el pago del recibo. Variable de naturaleza dicotómica que tomará valor 1 si el proceso administrativo se realiza de forma semestral y 0 en caso contrario.
- **Número de siniestros anteriores**: Número de siniestros anteriores incorporados a la póliza.
- **Edad del conductor asegurado:** Edad (en términos absolutos) del conductor asegurado en el instante de producirse el siniestro.
- **Coincidencia de apellidos:** Variable de naturaleza dicotómica. Alcanza valor 1 si existe concomitancia de apellidos entre la parte asegurada (tomador/propietario/conductor/ocupante) y los apellidos de la parte contraria. Toma valor 0 en caso contrario.
- **Día de la semana:** Variable de naturaleza dicotómica. Si el siniestro ocurre el fin de semana (sábado o domingo) tomará valor 1. En caso contrario tomará valor 0.
- **Diferencia de fechas:** Variable de naturaleza dicotómica. Tomará valor 1 si la fecha del accidente coincide con la fecha de efecto de la póliza con un margen de dos días. En caso contrario, tomará valor 0. La utilización de un margen de dos días incide en la posible ocurrencia de un siniestro sin tener el seguro en vigor por parte del tomador.
- **Provincia:** Indica la provincia donde se produce el siniestro.
- **Antigüedad de comunicación:** Antigüedad de notificación del siniestro en días. El cálculo es obtenido de restar la fecha de comunicación menos la fecha de ocurrencia del siniestro.
- **Variable objetivo:** La variable dependiente que pretendemos estimar. Esta variable objetivo indica si el siniestro analizado tiene indicios de fraude y/o sospechas de serlo o si, por el contrario, es legítimo.

Tabla 1. Codificación de las variables. Fuente: Elaboración propia.

Variable	**Código**
Canal de contratación	Tipoag
Antigüedad de la póliza	(Antigpol)
Variaciones del contrato	SUPLE
Forma de pago	Fop
Número de siniestros anteriores	His
Edad del asegurado	Edad
Coincidencia de apellidos	Destaf
Día de la semana	Sab
Diferencia de fechas	Dacc
Provincia	Proac
Antigüedad de comunicación	Anticom
Variable objetivo	Objetivo

Gráfico 1. Distribución de las variables objeto de análisis con RStudio

4. Selección de variables

Si la base de datos presenta excesivo ruido implicaría entumecer todo el proceso de clasificación o presentar resultados inexactos. De este modo, una adecuada selección de variables permitiría aminorar la dificultad del modelo, facilitando su correcta comprensión y acrecentando la precisión de los algoritmos si se distingue un conjunto apropiado. Es usual que la utilización de determinadas variables que conforman un subconjunto suscite mejores resultados que la utilización de todas las variables.

En el presente trabajo, se realiza una selección de las variables mediante el algoritmo Boruta desarrollada por Kursa y Rudnicki (2010). Este proceso de selección de variables se realiza de forma independiente del empleo de las metodologías de aprendizaje automático a implementar. La articulación del algoritmo se muestra a continuación:

1. Incorporación de aleatoriedad al conjunto de datos mediante la producción de variables sombra a partir de los predictores.
2. Entrenamiento del conjunto de datos transformados mediante un Random Forest (metodología subyacente) obteniendo una media de la importancia de las variables.
3. Constatación en cada iteración de si la variable real tiene mayor significación que la variable sombra.
4. El algoritmo concluye cuando las variables se aceptan o se rehúsan. Si se descubre que una variable original está por debajo de una variable sombra será una muestra de que su aportación al modelo será incierta y, por tanto, debe ser excluida.

El algoritmo Boruta busca apresar todas las características relevantes que puede tener un *dataset* respecto a una variable objetivo.

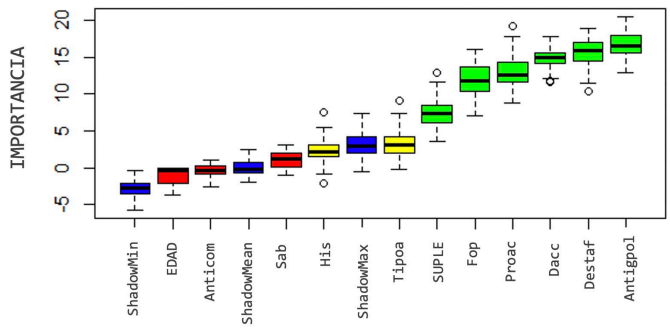

Gráfico 2. Análisis importancia de las variables.
Fuente: Elaboración propia, gráfico obtenido con RStudio.

La representación gráfica a través de un gráfico boxplots nos determina qué variables deberíamos considerar en el modelo (color verde), en color amarillo las variables cuya incorporación la valorará el experto, en color rojo las variables que podemos no tomar en consideración en el modelo y, para finalizar, en color azul son las denominadas variables sombra. El eje representa la importancia de cada variable dentro del conjunto de datos considerado.

Las variables que se introducen en el conjunto final para el entrenamiento de los distintos modelos son las siguientes:

- **6 variables relevantes**: Antigüedad de la póliza, fecha del accidente cercana a la contratación de la póliza, coincidencia de apellidos, forma de pago, provincia donde se produce el siniestro y el número de actas que tiene la póliza desde su emisión.
- **3 confirmadas como irrelevantes**: Edad, número de días que tarda en comunicar y si el siniestro se produce en fin de semana.
- **2 calificadas como tentativas**: siniestros asociados a la póliza anteriores al periodo estudiado y si se contrata a través de agente o no.

5. División del conjunto de datos

Una de las fases más significativas cuando elaboramos un pronóstico es la que se conoce como calibración (Silver, 2014). Los algoritmos de aprendizaje automático a desarrollar deben presentar una amplia capacidad de predicción, en caso contrario, sólo explicarían hechos ocurridos en el pasado reciente. El objetivo de la modelización es poder preludiar qué puede ocurrir bajo un umbral de probabilidad.

Elegido el conjunto de variables que constituye parte del análisis, se realiza un proceso de adquisición de información de los datos mediante la división de los datos en dos subconjuntos: El primero de entrenamiento y el segundo de constatación o test. Como se ha detallado, hemos de tener en cuenta que en base a los algoritmos elegidos puede acaecer que los modelos interpretan muy bien los datos de entrenamiento (nuestro objetivo no es la descripción) pero con escasa capacidad predictiva. De este modo, la división realizada contempla un 70% como datos del conjunto de entrenamiento y un 30% restante para la subsiguiente verificación de la fiabilidad del modelo propuesto. La capacidad predictiva se determinará al confrontar los

resultados de los verdaderos siniestros fraudulentos con los distintos pronósticos del modelo.

La división del conjunto de datos puede evidenciar una alteración del conjunto de datos de trabajo por un infraajuste o por un proceso de exceso. En nuestro caso, al realizar la división en estos dos conjuntos de datos, se observa como no se desvirtúa la muestra, obteniéndose prácticamente los mismos porcentajes de casos fraudulentos en todos los conjuntos de datos como se muestra en la siguiente **Tabla 2:**

Tabla 2. Proporción de casos fraudulentos. Fuente: Elaboración propia

	No Fraude	Fraude
Total datos	0.5643	0.4357
Entrenamiento	0.5657	0.4342
Test	0.5609	0.4391

6. Técnicas de Machine Learning para la detección del fraude

Las distintas metodologías de aprendizaje automático se vienen aplicando en el tratamiento de problemas de diversa índole. El problema de este trabajo es la detección de los siniestros fraudulentos mediante el empleo de un proceso automático que permita clasificar de forma recurrente operaciones sospechosas.

Se analizan distintos algoritmos pudiendo dividir en dos vertientes las herramientas y metodologías estadísticas utilizadas para la detección, tratamiento y prevención del fraude en el seguro de automóviles. En primer lugar, examinamos las técnicas clásicas basadas en la estadística tradicional, particularmente la regresión logística demostrando su eficacia en la detección del fraude a lo largo de los años, tal y como indican Bolton y Hand (2002). Son modelos utilizados para comprender la concomitancia entre una variable dicotómica binaria (fraude versus no fraude) con una o más variables explicativas independientes que pueden ser cualitativas o cuantitativas y estimar la probabilidad de que un siniestro sea fraudulento y/o sospechoso. **La regresión logística (**Logit**)** se engloba dentro de los Modelos Lineales Generalizados que utilizan como enlace la función logit. En otras palabras, permite catalogar siniestros dentro de las categorías de la

variable dependiente según la probabilidad que tenga de pertenecer a una de ellas en función de los valores que toman las covariables.

Adicionalmente, se emplea las **Máquinas de Vector Soporte** (Ben-Hur, Horn, Siegelmann & Vapnik, 2001) (SVM. Support Vector Machine) consistente en la utilización de una regresión lineal para estimar una función de decisión usando límite de clase no lineal basados en vectores de soporte. Las ventaja principal de este modelo es su simplicidad, puesto que únicamente quedan dos parámetros libres para ser elegidos, proporcionando una solución única, óptima y global si el entrenamiento del SVM se ha realizado para resolver un problema cuadrático linealmente limitado. Además, está basado en el principio de minimización del error total sobre el conjunto de datos, por lo que se consigue acotar el límite superior del riesgo real.

Otra técnica clásica muy utilizada en el aprendizaje automático son los **Arboles de Decisión** (DT: Decision Tree). Se determinan como un cúmulo de condiciones sistematizadas en un sistema jerárquico, de tal forma que la resolución final a tomar se decide siguiendo las directrices que se satisfacen desde la raíz del árbol hasta su hoja. De este modo, el árbol de decisión se transcribe según reglas y, por lo tanto, es sencillo de entender (Shmueli, Patel & Bruce, 2011). Asimismo, permite su representación gráfica, lo que faculta interpretar el modelo y realizar la conversión en reglas de decisión las hojas y las ramas del árbol de decisión.

Igualmente, los árboles de decisión se pueden emplear tanto para problemas de regresión, estimando el valor de la variable dependiente, como para problemas de clasificación mostrando una probabilidad de corresponder a una determinada clase. Se puede glosar como una serie de estipulaciones sucesivas partiendo del nodo raíz con el total de registros, que dan lugar a diferentes ramas, una por cada valor utilizable del atributo. Si el valor del atributo en un subconjunto tiene la análoga clase se convierte en hoja, en caso contrario será un nuevo nodo el que tendrá que proseguir discriminando. Los árboles de decisión particionan los datos de forma recursiva hasta que se satisface cierta condición como la minimización de la entropía o el encasillado de todos los registros. Por esta razón, la propensión es suscitar árboles y hojas con muchos nodos, implicando en algunas ocasiones un sobreajuste del modelo.

También se analiza el algoritmo **Naïve Bayes** (NB) y **las Redes Neuronales** (Hopfield, 1982) (NNET). En el primer caso, el modelo está fundamentado en la independencia de las variables predictores concediendo un alto grado

de precisión en las estimaciones. La metodología Naïve Bayes emerge por primera vez en la literatura del aprendizaje automático a finales de los años ochenta (Cestnik, Kononenko & Bratko, 1987).

En el segundo caso, tratamos de porfiar la estructura neuronal del cerebro humano. Así, las neuronas configuran capas con la siguiente estructura. Una primera capa donde se recibe directamente la información derivadas de las fuentes externas de la red, distintas capas ocultas y una capa de salida que transmite y traslada la información de la red hacia el exterior. Este proceso se puede automatizar para que sea el propio modelo el que se acomode a los datos según los rendimientos alcanzados.

A las técnicas precursoras, se integran las modernas técnicas de **bosques aleatorios** (RF: Random Forest) y **Extreme Gradient Boosting** (XGB) implantado para la detección de empresas susceptibles de realizar blanqueo de capitales (Badal-Valero, Álvarez-Jareño & Pavía, 2018). En el primer caso, la evolución congénita de los árboles de decisión son los bosques aleatorios o Random Forest (Breiman, 2001). Un bosque aleatorio es un clasificador consistente en una recopilación de clasificadores de árboles que son suscitados por un vector aleatorio distribuido idéntica e independientemente donde cada árbol acuña un voto para la clase más popular de entrada y los registros son encasillados con la clase que obtiene un mayor número de votos de los árboles que constituyen el ensamblado. La maniobra de los bosques aleatorios es compendiar de forma aleatoria un subconjunto de predictores para cimentar árboles de decisión donde cada árbol se despliega sobre una muestra bootstrap del cúmulo de entrenamiento. Como cada árbol depende de los valores de una muestra aleatoria de variables y el bosque aleatorio estándar es una combinación de las predicciones de los árboles con la misma distribución de todos los árboles en el bosque (Breiman, 2001), el bosque aleatorio no trabaja correctamente con conjuntos de datos que están muy desequilibrados como la localización de los siniestros fraudulentos en el seguro del automóvil. El error de los bosques aleatorios reconoce dos elementos especialmente: la correlación a través de los árboles del ensamblado y la existencia de cada árbol de manera particular.

El **bagging** incrementa la consistencia del árbol de decisión que junto a la elección aleatoria de variables acentúa la reciedumbre del modelo sobre la consideración de variables redundantes, haciéndolo muy apropiado en conjuntos de datos con abundantes variables.

En el segundo caso, el algoritmo Extreme Gradient Boosting (Friedman, 2001) es un ensamble de métodos de aprendizaje que encaja la potencia del pronóstico de diferentes metodologías en un único modelo yuxtapuesto. Es un algoritmo adaptativo en el que cada clasificador se erige en base a los outputs conseguidos en los clasificadores anteriores mediante la asignación de pesos a cada uno de los casos del conjunto de entrenamiento. Como explica Hidalgo (2014), el Extreme Gradient Boosting se determinó en su génesis como un método encaminado a reducir de forma ostensible el error de cualquier algoritmo, considerando el error menor que el clasificador aleatorio. Esta metodología ha estado preponderante en los últimos años en las competiciones y *challenges* de datos no sólo por su exactitud y concreción en la clasificación de instancias sino por la celeridad del procesamiento de la información. En este sentido, XGBoost (Chen et al., 2018) muestra una estructura de bloques en paralelo y que puede hacer uso de los diversos núcleos de la CPU así como el empleo eficiente del hardware.

7. Medidas para la evaluación de los modelos

Lo más usual es no poder cimentar modelos perfectos que clasifiquen de forma correcta todos los datos del conjunto de verificación y, por tanto, se deberá seleccionar la metodología que mejor trabaje la detección de los siniestros fraudulentos (Burez & Van den Poel, 2009).

La matriz de confusión es apropiada en el aprendizaje automático para cotejar los valores reales con los pronósticos en un modelo basal con datos entrenados. De este modo, se pueden pergeñar cuatro clases diferenciadas entre sí: Verdaderos negativos, verdaderos positivos, falsos negativos y falsos positivos.

El interés de la matriz reside en que no solo nos indica qué registros están correctamente catalogados, sino donde están encasillados aquellos que lo han sido incorrectamente. Kohavi y Provost (1998) exponen la matriz de confusión de acuerdo con la **Tabla 3** que se muestra a continuación:

Tabla 3. Matriz de Confusión Fuente: Elaboración Propia

		Predicción	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

Indagar la matriz de confusión una vez realizados los pronósticos puede descubrir algo exiguo a la hora de medir la eficiencia de un modelo de aprendizaje automático. La inmensa mayoría de los algoritmos de aprendizaje automático calculan la exactitud con la obtención de observaciones clasificadas de forma correcta. En este sentido, tenemos las siguientes métricas:

- TP es la tasa de verdaderos positivos para la categoría correspondiente y se obtiene dividiendo el resultado de las instancias correctamente clasificadas entre el total de instancias de esa categoría.
- FP: Tasa de falsos positivos. Conocidos como Errores Tipo I reconociendo aquellos expedientes como fraudulentos cuando realmente no lo son.
- TN: Tasa de casos negativos que se han clasificado de forma correcta.
- FN: Tasa de falsos positivos como la proporción de casos positivos correctamente clasificados del total de casos positivos existentes en el conjunto de verificación. Se conocen como Errores Tipo II.

De la matriz se pueden establecer diferentes medidas relativas a la precisión de los pronósticos, siendo la métrica más comúnmente utilizada la que hace referencia a la precisión y exactitud global del modelo (Accuracy). Nos indica la exactitud de cada clase y viene definida por los componentes de la diagonal principal dividido por el número total de registros donde se establece la relación entre el número de pronósticos correctos y el número total de predicciones. De la matriz de confusión se podrán alcanzar otra serie de indicadores. También establece el porcentaje de casos positivos predichos correctamente mediante el cociente entre los casos positivos correctamente clasificados del total de positivos pronosticados por el modelo (Keramati et al., 2014). También destaca la sensibilidad que medirá cuantos registros de la categoría positiva se etiquetan correctamente.

El Estadístico Kappa es un índice que coteja la exactitud del modelo con el alcanzado de forma aleatoria especificado como la disimilitud entre la precisión general y la precisión esperada dividida por 1 menos la precisión esperada. El estadístico presenta valores entre 0 y 1, obteniéndose el valor 1 cuando pronostica perfectamente el output (Kaymak, Ben-David y Potharst, 2012). La curva ROC y el área bajo la curva ROC (AUC) se emplean para evaluar la precisión de una clasificación. Esta curva ROC evalúa el rendimiento de un modelo de clasificación entre las tasas de positivos

correctamente clasificados (TP) y la tasa de negativos incorrectamente clasificados como positivos (FP) (Swets, 1988). Los valores fluctuarán entre 0,5 y 1 donde a mayor área bajo la curva mejor será la exactitud del algoritmo donde los valores cercanos a 1 indicarían un buen ajuste del modelo en relación al conjunto de datos.

8. Resultados

Huigevoort (2015) expone que con las métricas trazadas anteriormente se podrá elegir el mejor modelo para cada una de las distintas técnicas y luego confrontarlos con otras metodologías. Así, el árbol de decisión se realiza mediante el algoritmo *rpart* (Therneau & Atkinson, 2018) de R, donde el algoritmo sirve para encasillar los registros del conjunto de entrenamiento mediante árboles de decisión. Para la regresión logística se emplea la librería *MASS* de R con la función *glm* mediante la utilización de la función enlace logit.

La metodología Random Forest es entendida como un ensamble de árboles de decisión y se realiza mediante la librería *randomForest* (Liaw & Wiener, 2002) de R. De este modo, se seleccionan aleatoriamente un número de variables y se construye un árbol. Este procedimiento se reitera un número grande de veces y se establece el resultado del modelo ensamblado mediante votación. Una instancia será clasificada como fraude cuando la mayoría de los árboles la hayan clasificado de esta forma. No obstante, los resultados de los bosques aleatorios son arduos de explicar ya que son la consecuencia de la anexión de diferentes árboles de decisión.

La metodología SVM se cimienta en la minimización del riesgo estructural, de modo que se busca construir modelos que tengan poco riesgo de cometer errores ante distribuciones futuras. Dicho método sirve para resolver el problema de clasificación en siniestros fraudulentos en los que a partir de unos dato de entrenamiento se construye un hiperplano capaz de dividir los puntos en dos grupos. En R se disponen de diferentes paquetes con los que implementar SVM, sin embargo, el paquete recomendado por Karatzoglou, Meyer y Hornik (2006) es *svmpath* desarrollado por Hastie, Rosset, Tibshirani y Zhu (2004).

El algoritmo Extreme Gradient Boosting, como el caso precedente, es un ensamble de árboles de decisión pero adaptando una metodología distinta a la del Random Forest. Así, el boosting es analizado como un algoritmo de optimización mediante la elección iterativa de una función que apunta en la dirección del gradiente negativo (Breiman, 2001). En el algoritmo *xgboost*

de R (Chen et al., 2018) se utilizarán árboles de decisión y la función a optimizar será el área bajo la curva ROC.

A continuación, se determinan los resultados obtenidos con los diferentes modelos, así como las distintas conclusiones conseguidas:

Tabla 4. Resultados de los modelos Fuente: Elaboración propia.

	Logit	SVM	NB	NNET	DT. Decisión Tree	RF. Random Forest	XGB
Accuracy	0,6699	0,5962	0,6090	0,6635	0,5929	0,6058	0,6026
Kappa	0,3097	0,1445	0,1449	0,2975	0,1575	0,1756	0,1837
TP	0,8011	0,7784	0,8920	0,7898	0,6989	0,7443	0,6818
TN	0,5000	0,3603	0,2426	0,5000	0,4559	0,4265	0,5000
AUC	0,6502	0,5694	0,5673	0,6449	0,5774	0,5854	0,5909

En relación a la idoneidad representativa de los modelos, estos no presentan diferencias fundamentales y todos obtienen una precisión global del modelo similar. En relación a la capacidad predictiva, el modelo Logit es el que mejor predice en términos generales. De este modo, consigue la mayor precisión (66,99%), el mayor estadístico Kappa, el mayor porcentaje de acierto en los verdaderos negativos y mayor área bajo la curva ROC (AUC). El objetivo es la obtención del mejor pronóstico de los verdaderos positivos y un soporte en la toma de decisiones en cuanto a la tramitación de los siniestros, mejorando la cultura de análisis y toma de decisiones basada en los datos, tomando en consideración la mejora en el control y la gestión de los riesgos bajo la normativa de Solvencia II.

No obstante, la optimización del umbral de probabilidad a partir del cual un expediente es catalogado como fraude puede ayudar a mejorar el porcentaje de verdaderos positivos.

9. Conclusiones

El fraude en el seguro del automóvil no es un fenómeno estático, y los algoritmos para su prevención y detección deberían contemplar la volatilidad del comportamiento del fraude mediante un re-entreno frecuente de los distintos modelos, así como la introducción de modelos de respuesta

incremental para tratar las nuevas situaciones mediante la inclusión de un grupo de control para su correcto análisis.

La detección del fraude en el seguro del automóvil permite diferentes metodologías estadísticas para su correcto tratamiento a lo largo del tiempo. Representando una herramienta que apoye al trabajo del tramitador y oriente los limitados recursos económicos destinados a la detección y prevención del fraude a determinados siniestros con mayor probabilidad de serlo. La aplicación de nuevas técnicas de detección no pretende sustituir la labor del tramitador, sino facilitar su labor teniendo en cuenta que el fraude deteriora el resultado técnico y la siniestralidad en plena guerra de precios del sector.

El establecimiento y determinación de una matriz de confusión permite analizar qué siniestros legítimos han sido clasificados como fraude y qué siniestros fraudulentos han sido incorrectamente clasificados en cualquiera de los dos sentidos. Es fundamental enriquecer la predicción de los modelos mediante la incorporación de nuevos casos y para evitar el uso ineficiente de los recursos en la tramitación de siniestros cuya demostración final estará en entredicho.

En estos casos, la solución final deberá ser analizada por la entidad aseguradora en base al proceso establecido para la prevención y detección del fraude. La solución tecnológica a considerar dependerá de política general de la empresa. Desde el momento en el que se conocen las probabilidades de fraude de un siniestro será necesario fijar qué expedientes son merecedores de una investigación más minuciosa y analizar cuáles serán las acciones a realizar en base a los resultados obtenidos.

No obstante, sería interesante abordar el planteamiento del problema (detección y prevención del fraude) bajo el prisma de la tipología de siniestros. Todo ello mejoraría la capacidad predictiva de los modelos al incluir variables complementarias a las utilizadas en este trabajo. Esta personalización para cada tipología de siniestro, permitirá conocer cuáles son las variables que tienen una mayor ponderación para cada asegurado. La personalización de la estrategia a seguir permite incrementar la eficiencia del departamento de siniestros y/o fraude.

Aunque en este trabajo se ha implementado un muestreo de los datos para el balanceo de las clases a través de la adición o eliminación de registros, es necesario seguir profundizando en el tratamiento del problema de datos desequilibrados para el conjunto de una cartera de siniestros del ramo de automóviles. Así, Chen, Liaw y Breiman (2004) proponen el balanceado mediante la generación sintética de datos y el aprendizaje sensible a costes.

Y Yen y Lee (2009) infieren que la generación de datos sintéticos aplica mejor que el método de sobremuestreo aleatorio y elude el sobreajuste. También sería conveniente considerar otros métodos avanzados para balancear los conjuntos de datos desequilibrados como, por ejemplo, el basado en clústers, métodos basados en kernel y la utilización de muestreo sintético adaptativo.

Finalmente, es imprescindible que las entidades aseguradoras sean capaces de recopilar progresivamente información de sus asegurados en diferentes ámbitos para mejorar el conocimiento del colectivo, y siempre bajo el cumplimiento estricto de la normativa RGPD de protección de los datos.

10. Bibliografía

Artís, M., Ayuso, M., Guillen, M. (1999). Técnicas cuantitativas para la detección del fraude en el seguro del automóvil. *Anales del Instituto de Actuarios Españoles*, 5, 51-84.

ASEPEG (2020). *Glorario de Términos*. En https://www.apeseg.org.pe/glosario-de-terminos/

Ayuso, M., Guillén, M. (1999). Modelos de deteccion de fraude en el seguro de automóvil, *Cuadernos Actuariales*, 8, 135-149.

Badal-Valero E., Alvarez-Jareño, J.A. y Pavía, J.M. (2018). Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish court case, *Forensic Science International*, 282, 24-34.

Belhadji, B., Dionne, G. (1997). Development of an expert system for automatic detection of automobile insurance fraud. Working Paper 97-06. *École des Hautes Études Commerciales*. Université de Montréal.

Ben-Hur, A., Horn, D., Siegelmann, H., Vapnik, V. (2001). Support Vector Clustering. *Journal of Machine Learning Research*. 2. 125-137.

Bolton, R.J. y Hand, D.J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, 17 (13), 235-255.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1), 5-32.

Brockett, P.L, Xia, X y Derrig, R. (1995).Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. *Journal of Risk and Insurance*, 65 (2), 245-274.

Burez, J. y Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Experts Systems with Applications*, 36, 4626-4636.

Cestnik, B, Kononenko, I, Bratko, I. (1987). A knowledge elicitation tool for sophisticated users. *Progress in Machine Learning*, 31-45, Sigma Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-*sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.*

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 853-867. Springer US.

Chen, C.; Liaw, A. y Breiman, L. (2004). *Using random forest to learn imbalanced data*. Technical Report 666. Statistics Department of University of California at Berkley.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H.; Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M.,L Geng, Y. y Li, Y. (2018). xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgboost

Crocker, K.J. y S. Tennyson (2002). Insurance Fraud and Optimal Claims Settlement Strategies. *Journal of Law & Economics*, 45(2), 469-507.

Cummins, J.D. y Tennyson, S. (1996). Moral Hazard in Insurance Claiming: Evidence from Automobile Insurance. *Journal of Risk and Uncertainty*, 12 (1), 29-50.

Derrig, R.A y Ostaszewski, K.M. (1995). Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification, *Journal of Risk and Insurance*, 62 (3), 447-482.

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

Guo, H., Li, Y., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: *Review of methods and applications.Expert Syst*. Appl., 73:220–239.

Hastie T, Rosset S, Tibshirani R, Zhu J (2004). The Entire Regularization Path for the Support Vector Machine. *Journal of Machine Learning Research*, 5, 1391–1415.

He, H. y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

Hidalgo Ruiz-Capillas, S (2014). *Random Forests para detección de fraude en medios de pago*. Trabajo Final de Máster. Universidad Autónoma de Madrid.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79 (8), 2554-2558.

Huigevoort, Chantine (2015). *Customer Churn prediction for an insurance company*. Eindhoven University of Technology. Master Thesis. https://pure.tue.nl/ws/portalfiles/portal/47019808 [Último acceso: 23 de septiembre de 2018]

ICEA (2018). *El Fraude al Seguro Español. Estadística a diciembre. Año 2017*. Madrid, España.

Kaymak, U.; Ben-David, A. y Potharst, R. (2012). The AUK: A simple alternative to the AUC. *Engineering Applications of Artificial Intelligence*, 25 (5), pp. 1082-1089.

Karatzoglou, A., Meyer, D. y Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, 15 (i09).

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., y Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24, pp. 994-1012.

Kohavi, R., y F. Provost (1998) On Applied Research in Machine Learning. In Editorial for the *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, Columbia. University, New York, 30.

Kursa, M.B. y Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 1-13. URL: http://www.jstatsoft.org/v36/i11/

Liaw, A. y Wiener M. (2002). Classification and Regression by Random Forest. *R News* 2 (3), pp 18-22.

Lunardon, N., Menardi, G. y Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *The R Journal*, 6 (1), 82-92.

López, V., Fernández, A., García, S. Palade, V. y Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*. 250, 113-141.

Picard, P. (2000). *Economic analysis of insurance fraud. Handbook of insurance*. 315-362. Springer, Dordrecht.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M. y Baesesns, B. (2017). Network-based Fraud Detection for Social Security Fraud. *Management Science*, 63 (9), 3090-3110.

Shmueli, G.; Patel, N.R. y Bruce, P.C. (2011). *Data mining for business intelligence: concepts, techniques, and applications in microsoft office excel with xlminer*. John Wiley and Sons, second edition.

Silver, N. (2014). *La Señal y el Ruido*. Ediciones Península, Barcelona.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Scienc*e, 240, 1285-1293.

Therneau, T. y Atkinson, B. (2018). rpart: *Recursive Partitioning and Regression Trees*. R package version 4.1-13. URL: https://CRAN.R-project.org/package=rpart

Yen, S.J y Lee, Y.S (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36 (3), 5718-5727.