Análisis del desarrollo de modelos de Machine Learning para modelos de clasificación y predicción en la distribución de seguros

Proyecto ABMS 3.0

Enero de 2024

codeoscopic

Estudio auspiciado por:

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

Financiado por la Unión Europea

Plan de Recuperación, Transformación y Resiliencia

Contenido

Análisis del desarrollo de modelos de Machine Learning para modelos de clasificación y predicción en la distribución de seguros 1
Proyecto ABMS 1
Introducción 3
Análisis desarrollados 3
1 Clasificación o clusterización de clientes 3
2 Elaboración de un modelo de propensión a la baja 4
3 Elaboración de un modelo de propensión al fraude 4
Número de casos 4
1 MODELO DE CLASIFICACIÓN 5
Metodología 5
Clasificaciones 5
Resultados obtenidos 7
2 MODELO DE PROPENSIÓN A LA BAJA 10
Metodología 10
Creación de modelos de aprendizaje 10
Resultados obtenidos 11
Correlaciones entre variables 11
Resultados de acierto con modelos de aprendizaje automático K-NN y RF 12
Porcentaje de la cartera de clientes es propensa a la baja 13
3 MODELO DE DETECCIÓN DE FRAUDE 14
Metodología 14
Modelo 1: El modelo de declaración de siniestro desde fecha de alta de la póliza 14
Modelo 2 El modelo de número de pólizas 15
Modelo 3 Modelo de tiempo de pólizas 16
Modelo 4 Modelo de suma de primas 17
Modelo 5 Modelo de suma de primas con valor del siniestro 18
Resultados obtenidos 18
Modelo 1: El modelo de declaración de siniestro desde fecha de alta de la póliza 18
Modelo 2 El modelo de número de pólizas 19
Modelo 3 Modelo de tiempo de pólizas 20
Modelo 4 Modelo de suma de primas 21
Modelo 5 Modelo de suma de primas con valor del siniestro 22
Fichero de salida 22

Introducción

Este documento corresponde a una fase del proyecto que Codeoscopic ha puesto en marcha con el nombre de ABMS 3.0, cuyo objetivo es desarrollar IA para incluir elementos predictivos en sus soluciones que permitan agilizar las operaciones y servicios en las herramientas que la empresa pone a disposición de los mediadores de seguros.

El desarrollo de este proyecto está soportado por los fondos europeos de recuperación 'Next Generation EU' para proyectos de investigación y desarrollo en inteligencia artificial y otras tecnologías digitales y su integración en las cadenas de valor. Codeoscopic optó a las ayudas por medio de la convocatoria de Red.es.

El siguiente análisis recoge parte del conocimiento desarrollado en los procesos de desarrollo de modelos de clasificación y predicción sobre una base de datos de clientes de seguros con más de 34 millones de registros entre clientes, pólizas, recibos, siniestros y corredurías.

Parte de la investigación y desarrollo para el proyecto ABMS 3.0 se realiza con el Instituto Universitario de Investigación de Biocomputación y Física de Sistemas Complejos (BIFI) de la Universidad de Zaragoza, dentro de un acuerdo de colaboración alcanzado por Codeoscopic.

Análisis desarrollados

Los objetivos fijados por el proyecto son crear modelos de:

1. Clasificación o clusterización de clientes

Los modelos de clusterización de Inteligencia Artificial (IA) son un tipo de algoritmos de aprendizaje no supervisado que se utilizan para agrupar objetos o datos similares en grupos o conjuntos. Estos modelos se centran en identificar grupos de registros similares y en etiquetar registros según el grupo al que pertenecen. Esto se lleva a cabo sin la ventaja de disponer de conocimientos previos sobre los grupos y sus características.

Dentro del contexto de la distribución de seguros, el objetivo es encontrar agrupaciones de clientes sofisticadas con alto nivel predictivo en su conducta futura.

Esta clasificación puede servir para:

- Personalizar productos y satisfacer mejor las necesidades de cada grupo.
- Identificar oportunidades de venta cruzada y up-selling.
- Realizar una mejor gestión del riesgo al identificar grupos de clientes que presentan un mayor o menor riesgo.
- Mejorar la retención del cliente.
- Realizar campañas de marketing más efectivas.

Para el distribuidor de seguros, tener una mayor comprensión más profunda de sus clientes, lo que puede informar una amplia gama de decisiones estratégicas y operativas.

La posibilidad de poder determinar la fidelidad de clientes es útil para personalizar servicios, retener clientes, crear un marketing más efectivo y mejorar la satisfacción del cliente.

2. Elaboración de un modelo de propensión a la baja

Para la elaboración de modelos de propensión a la baja se ha desarrollado un modelo de machine learning de clasificación binaria. Este es un tipo de algoritmo de aprendizaje automático que se utiliza para clasificar los datos en dos grupos o categorías distintas. La "binaria", en la clasificación binaria, se refiere a estas dos categorías en las que los datos pueden ser clasificados como si será baja o no será baja.

Este modelo puede servir para retener clientes, entender sus necesidades, optimizar recursos al conocer cuáles están en riesgo de fuga y mejorar la satisfacción del cliente.

3. Elaboración de un modelo de propensión al fraude

Para la elaboración de modelos de propensión a la baja se han desarrollado cinco modelos de machine learning de clasificación de agrupaciones y cálculo de percentiles.

Número de casos

El estudio se ha realizado sobre un conjunto de datos que incluye información de clientes en el sector de seguros, tanto personas físicas como jurídicas. Se han utilizado variables como duración de pólizas, número de pólizas, prima media, número de compañías y cambios de compañía, entre otras. Además, se han segmentado los datos por tipo de persona (física o jurídica) para un análisis más detallado. Sin embargo, el número exacto de datos no ha sido proporcionado en las referencias proporcionadas.

Para la modelización se tomaron las tablas de:

Tabla	Registros
Cliente	1.377.505
Póliza	3.469.604
Recibos	17.575.338
Siniestro	1.885.501

1 MODELO DE CLASIFICACIÓN

Metodología

Se han probado distintos modelos de segmentación para poder clasificar a los clientes en distintos grupos. Debido a la naturaleza de los datos proporcionados, está división se tendrá que hacer mediante el empleo de los datos de las pólizas y las primas asociadas a cada cliente. Los algoritmos clasificadores pueden dividir a los clientes en grupos y analizando variables como el número de pólizas, duración de las pólizas, y número de compañías.

Entre los distintos algoritmos clasificadores se han obtenido los resultados más interesantes con el algoritmo K-Prototypes. Este algoritmo emplea K-Means para las variables numéricas y la distancia de Hamming para las categóricas.

Las variables utilizadas para entrenar el algoritmo de segmentación son las siguientes:

- Nº cambios compañía por NIF: número de cambios de compañías de un cliente en cada ramo en concreto.
- Días medios por póliza: duración media en días de las pólizas de cada cliente.
- Nº compañías por NIF: número de compañías en las que ha estado un cliente en total.
- Localización: los dos primeros dígitos del campo CPostal.
- Prima media por NIF: el valor medio de la prima por cliente.
- Nº pólizas por NIF: número de pólizas que tiene un cliente.
- Nº ramos por NIF: número de ramos en los que tiene pólizas un cliente.

Estas variables se utilizan para clasificar a los clientes en distintos grupos, lo que permite segmentarlos en función de su comportamiento en relación con las pólizas y las primas de seguros.

Clasificaciones

Los grupos encontrados a través del proceso de clustering representan distintos comportamientos y características de los clientes en el sector de seguros. En particular, se identificaron clústeres que se diferencian en función de variables como la duración media de las pólizas, el valor medio de la prima por cliente, el número de pólizas que tiene un cliente, y el número de ramos en los que tiene pólizas un cliente.

Clúster	Segmento	% Segmento	Nº pólizas por NIF	Nº compañías por NIF	Nº cambios Compania por NIF	Volumen prima	Días Medios por Póliza
0	Oportunistas	18%	2,8923	1,9698	0,3987	1.031,07 €	963
1	Mercenarios	8%	5,0896	3,1919	0,9751	1.867,94 €	956
2	Fieles de bajo consumo	14%	1,3009	1,0102	0,0050	451,46 €	2.648
3	Alto potencial	60%	1,3012	1,0000	0,0049	454,22 €	667

Una primera clasificación es identificar las variables que más discrimina los grupos. En este caso, el **Nº de pólizas por NIF**, que crea dos grupos, uno con los clúster 0 y 1, grupos de más de dos pólizas, y los clústers 2 y 3, que son grupos de menos de dos pólizas.

Una segunda clasificación es el de **Nº de cambios compañía por NIF**, en el que se encuentran nuevamente los clúster 0 y 1 con conductas de cambio, mientras que los clúster 2 y 3 apenas presentan cambios.

Así, se presentan dos grupos diferenciados, uno con muchas pólizas por cliente y cambios de compañía, y otro con pocas pólizas por cliente y sin apenas cambio de compañía.

Un análisis individualizado por cada cluster nos da la siguiente descripción:

Oportunistas

El clúster 0 se presenta como un cliente bien equipado, con una densidad de seguros de **2,9 pólizas**, un alto nivel de cambios de compañía. No es un cliente integral, porque hay algunos clientes que tienen un nivel más alto, ni tampoco un cliente que está en un número de pólizas bajo próximos al uno como ocurre con otros segmentos. Su nivel de antigüedad en la correduría es medio, situado en 963 días medios por póliza.

Este grupo constituye el **18% de la cartera de clientes**.

Mercenarios

El clúster 1 tiene unos valores próximos al segmento Oportunistas, pero con una mayor densidad de seguros con **5 pólizas**, que supone más de un 76% de seguros que el cliente oportunista, y también un mayor nivel de cambio de compañías, que supera en casi 1,5 veces la conducta de los oportunistas. Esto quiere decir que su nivel de oportunismos es mayor, y casi se podría decir que tiende más a una conducta mercenaria.

Este grupo constituye el **8% de la cartera de clientes**.

Fieles de bajo consumo

El clúster 2 se destaca por ser el cliente con mayor longevidad y con un nivel bajo de densidad, con **1,3 pólizas** de media. La antigüedad del cliente alcanza un máximo, con 2.648 días medios

por póliza y un nivel de cambio de compañía muy bajo. Estos clientes se podrían considerar como mono-pólizas y sin potencial.

Este grupo constituye el **14% de la cartera de clientes**.

Alto potencial

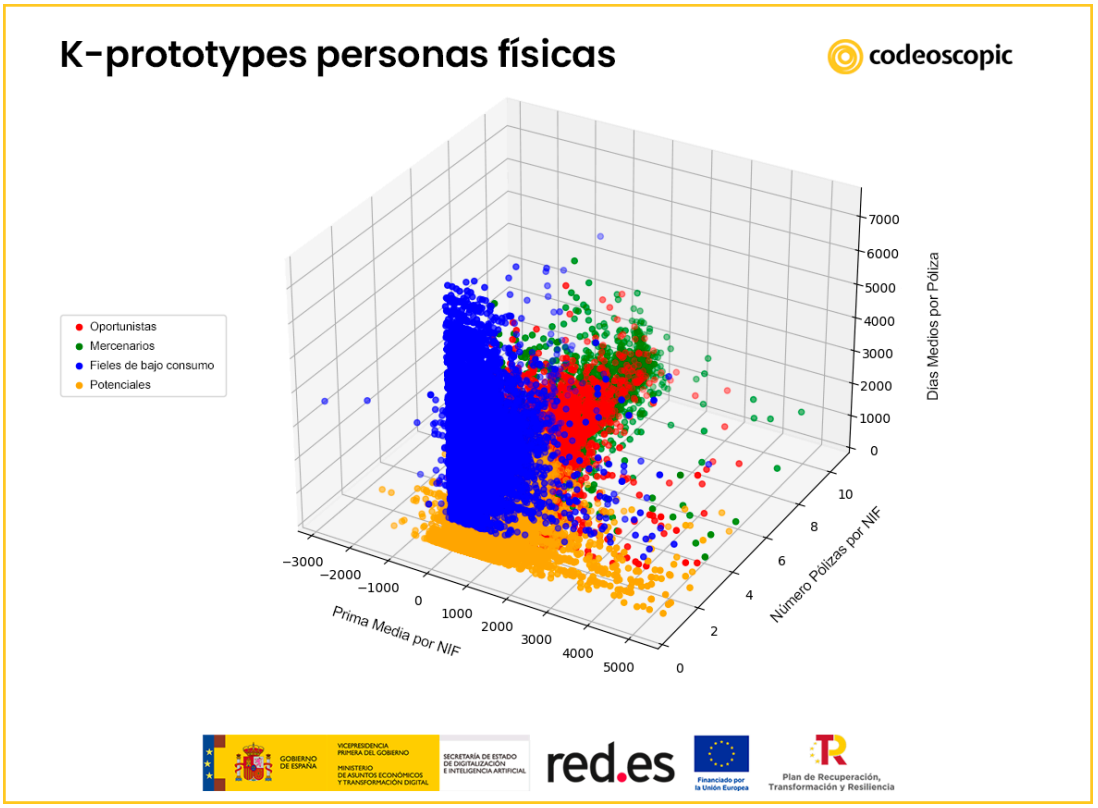
El clúster 3 está en sus valores muy próximo al cliente *Fiel de bajo consumo*, pero con la diferencia que lleva poco tiempo en la correduría, con una media con 667 días medios por póliza. Así, su media es de 1,3 pólizas por cliente, que coincide con los fieles de bajo consumo, y su nivel de cambio de compañía es muy bajo.

Este grupo constituye el **60% de la cartera de clientes**.

Resultados obtenidos

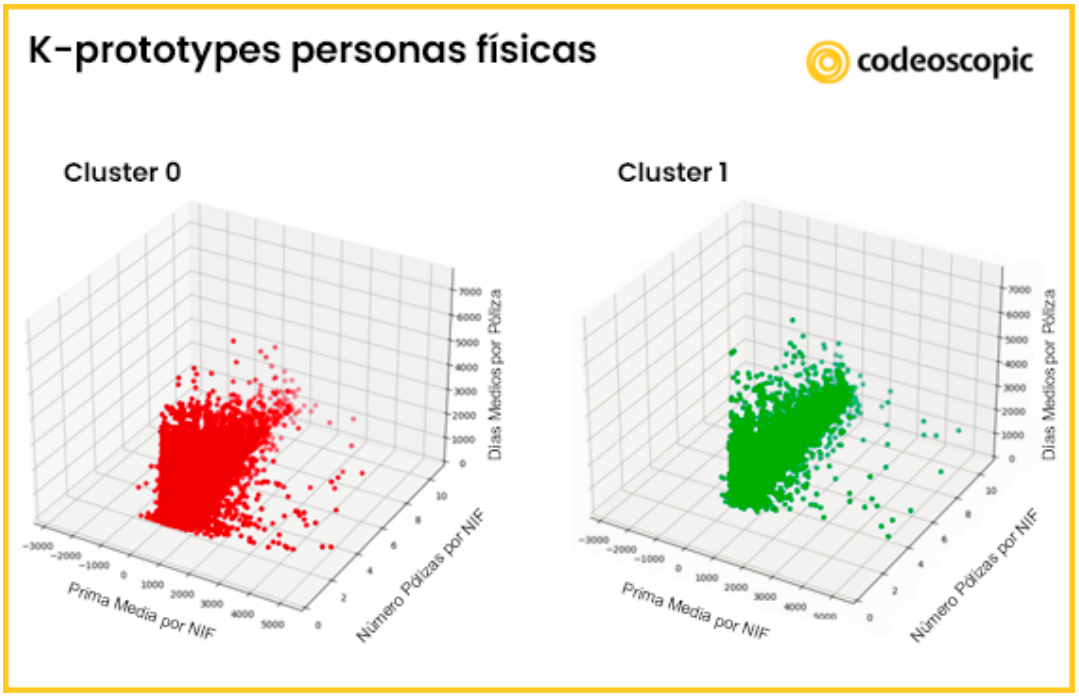
Para ver los resultados de forma visual, se pueden realizar representaciones de las variables 3 a 3 y separarlas por los clústeres obtenidos.

En la siguiente gráfica se muestran los clústeres hallados discriminados por Prima Media por NIF, el Nº Pólizas por NIF y Días Medios por Póliza.

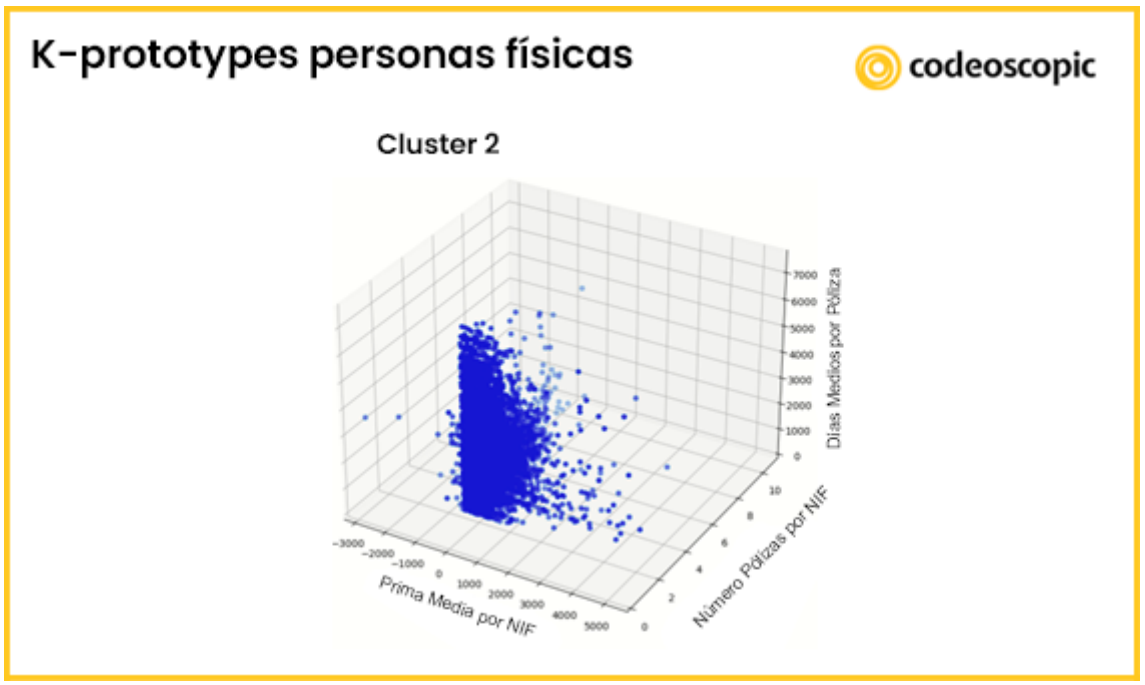


Si separamos en diferentes capas los cuatro clústeres hallados se pueden ver las diferencias y similitudes.

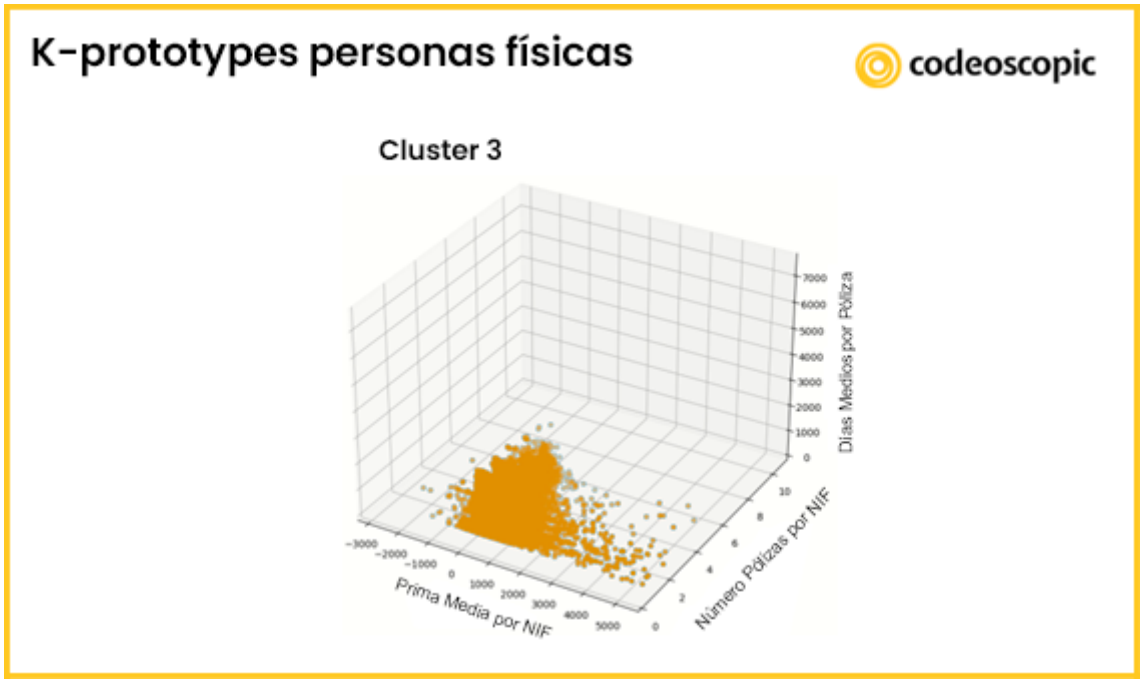
En la siguiente gráfica se muestran las similitudes que presentan los clústeres 0 y 1, con un alto valor de pólizas y conductas oportunistas y mercenarias. Visualmente se observa la prolongación de las observaciones a lo largo del eje nPólizasporNIF, con la diferencia de pólizas medias, por lo que se sitúa un clúster más adelante que el otro.



En la siguiente figura, vemos cómo destaca la conducta del clúster 2, con un elevado número de días de antigüedad media de la póliza, ascendiendo por el eje vertical que mide los días de media. Es el único segmento que posee esta configuración y que se diferencia de los demás, además de una fuerte concentración alrededor de entre una y dos pólizas.



Por último, podemos observar el clúster 3 con una baja densidad de pólizas y pocos días de media por póliza, que corresponde al eje vertical.



2 MODELO DE PROPENSIÓN A LA BAJA

Metodología

El modelo de propensión a la baja analiza la correlación de las variables utilizadas en los modelos para realizar un estudio sobre la capacidad predictiva de los modelos. Describe las nuevas características creadas a partir de los datos proporcionados y su correlación con la variable objetivo, las cancelaciones de pólizas.

Se utilizaron varios modelos de aprendizaje automático para predecir las anulaciones de pólizas, incluyendo árboles de decisión, bosques aleatorios, y modelos de regresión logística. Estos modelos fueron ajustados para abordar el desequilibrio de clases mediante técnicas como el ajuste de pesos de clase, la submuestreo de la clase mayoritaria y el sobremuestreo de la clase minoritaria. Estas técnicas fueron implementadas para mejorar la capacidad predictiva de los modelos y mitigar el impacto del desequilibrio de clases en la precisión de las predicciones.

Las variables utilizadas en la confección de los modelos han sido:

- El aumento de la prima en tanto por uno en el último año.
- Duración media de la póliza por cliente (por único NIF).
- Antigüedad de la póliza (años).
- Sexo.
- Estado civil.

Los modelos utilizados pertenecen al aprendizaje automático o machine learning, y han sido elaborados a partir de esquemas generales incorporados a la librería de Python scikit-learn. En términos generales, se trata de un problema de clasificación binaria (variable objetivo baja/no baja), donde se utilizan variables de tipo numérico y de tipo categórico.

Creación de modelos de aprendizaje

Tras el completo filtrado de los datos, disponemos de 234.991 datos. Existe un desbalanceo entre las clases a predecir, de forma que se tienen 152.895 valores catalogados como «No baja» (un 65 %), y 82 096 reconocidos como «Baja» (un 35 %).

Tras un breve ensayo con diferentes modelos de aprendizaje automático, se han seleccionado dos de ellos: el «algoritmo de los vecinos» o k-Nearest Neighbors (abreviado k-NN) y algoritmos basados en árboles de decisión, en especial, Random Forest (abreviado RF).

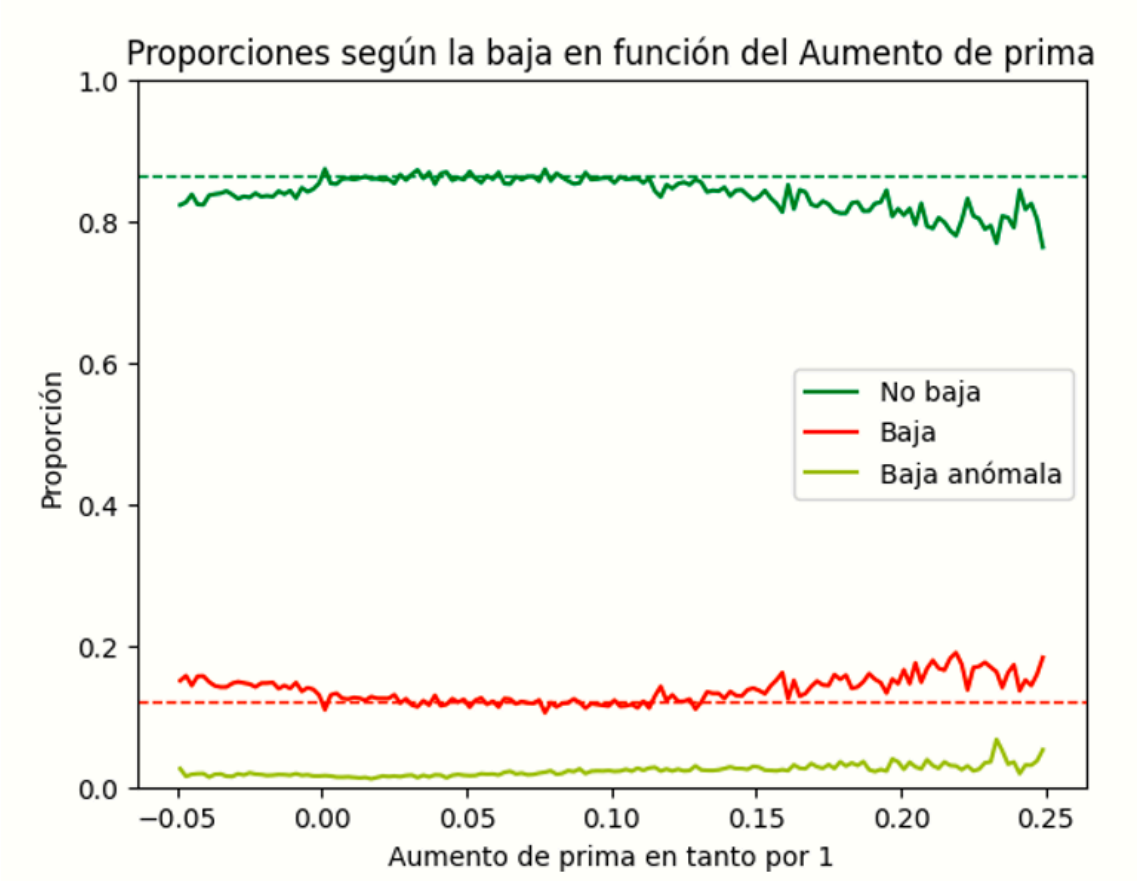
La evaluación del modelo se ha llevado a cabo siguiendo dos técnicas; una validación cruzada 10-fold y una separación entrenamiento - validación.

Para el criterio seguido para considerar un modelo aceptable se requiere una alta puntuación en ambos test, prediciendo con igual exactitud, aproximadamente, tanto la «Baja» como la «No baja».

Resultados obtenidos

Correlaciones entre variables

Hubo correlación entre los aumentos de primas y las anulaciones de pólizas. Se observó que un aumento de prima mayor se correspondía con una mayor proporción de anulaciones, especialmente cuando el aumento de prima supera un cierto límite. Esta correlación fue evidente en el análisis de las variables y se utilizó para predecir las cancelaciones de pólizas en el estudio.



En el gráfico se puede observar el comportamiento de las bajas. Aquí están representadas todas las pólizas que han visto incrementado su prima a lo largo del tiempo. Si los incrementos están por debajo del 12%, la proporción de bajas es constante. Si el incremento de bajas supera este porcentaje, la línea roja que representa las bajas comienza a aumentar acercándose a niveles del 20% de bajas.

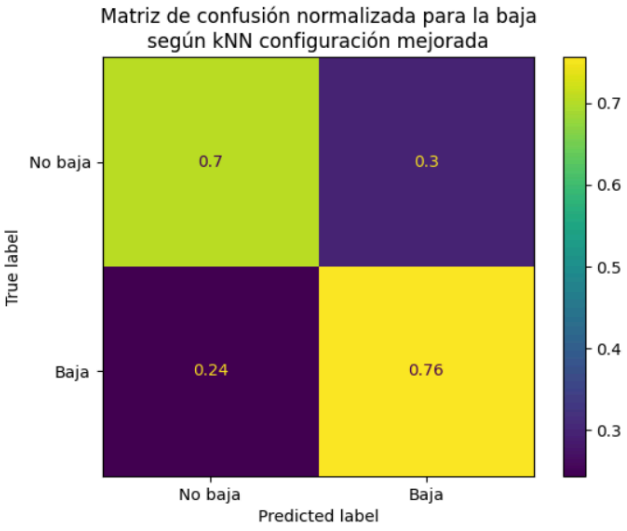
Además del aumento de precio, otras variables que tienen correlación con la baja del cliente incluyen el estado civil y el sexo del cliente. Se observó que los clientes con estado civil "Casado/as" tenían en general más bajas que los clientes con estado civil "Solteros/as", y que los clientes con sexo "Hombre" también eran más propensos a la baja que los clientes con sexo "Mujer".

Resultados de acierto con modelos de aprendizaje automático k-NN y RF

Los modelos base estudiados han sido k-Nearest Neighbors (k-NN) y Random Forest (RF) se comportan de forma similar, con un porcentaje de acierto de aproximadamente el 75%, acertando independientemente de si ha existido una baja o no (74.3 ± 2.4 % de precisión).

En el modelo k-NN observamos la siguiente figura se muestran los estudios de **acierto del modelo** con los siguientes correlatos:

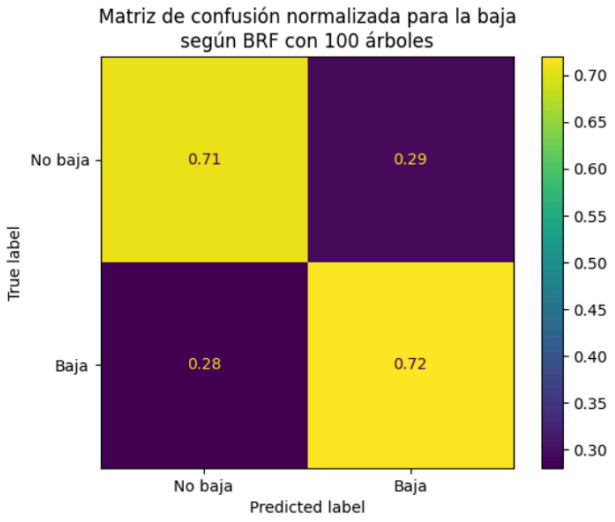
- 70% Acierto de No baja, en clientes reales de No baja.
- 76% Acierto de Baja en clientes reales de Baja.



Los **errores de modelo** con No acierto en la predicción de los casos son:

- 30% No acierto con predicción de Baja y en clientes reales era No baja.
- 24% No acierto con predicción de No baja y en clientes reales era Baja.

El segundo modelo de algoritmo utilizado es el de Random Forest o RF construye múltiples árboles de decisión durante el entrenamiento. Los resultados obtenidos son similares al anterior, aunque disminuye levemente en la posición de predicción de Baja que acierta realmente la Baja (72%), y sube un punto la posición de predicción No baja que acierta la No baja (71%).



Porcentaje de la cartera de clientes es propensa a la baja

El análisis de la base de datos completa determinó que son propensos a la baja el 20,5% de los clientes.

3 MODELO DE DETECCIÓN DE FRAUDE

Metodología

El apartado de modelos de detección de fraude propone cinco modelos los cuales se basan en el análisis de datos de siniestros, pólizas y primas de los clientes.

El modelo de alarmas de fraude analiza los clientes "outliers" para distintas casuísticas: con respecto al número de siniestros se compara con el número de pólizas, duración de las pólizas y suma de las primas de las pólizas. Con respecto al desembolso de los siniestros, se compara con la suma de las primas. Así se obtienen los clientes que se desvían mucho con respecto a los demás, siendo posible que este desviamiento se deba a algún comportamiento fraudulento. Se puede modificar el fichero de configuración para modificar los parámetros que hacen saltar las alarmas.

Cada modelo utiliza diferentes técnicas estadísticas para identificar comportamientos anómalos y sospechosos que podrían indicar la presencia de fraudes. Los modelos son configurables y se pueden ajustar según las necesidades de cada empresa aseguradora.

Modelo 1: El modelo de declaración de siniestro desde fecha de alta de la póliza

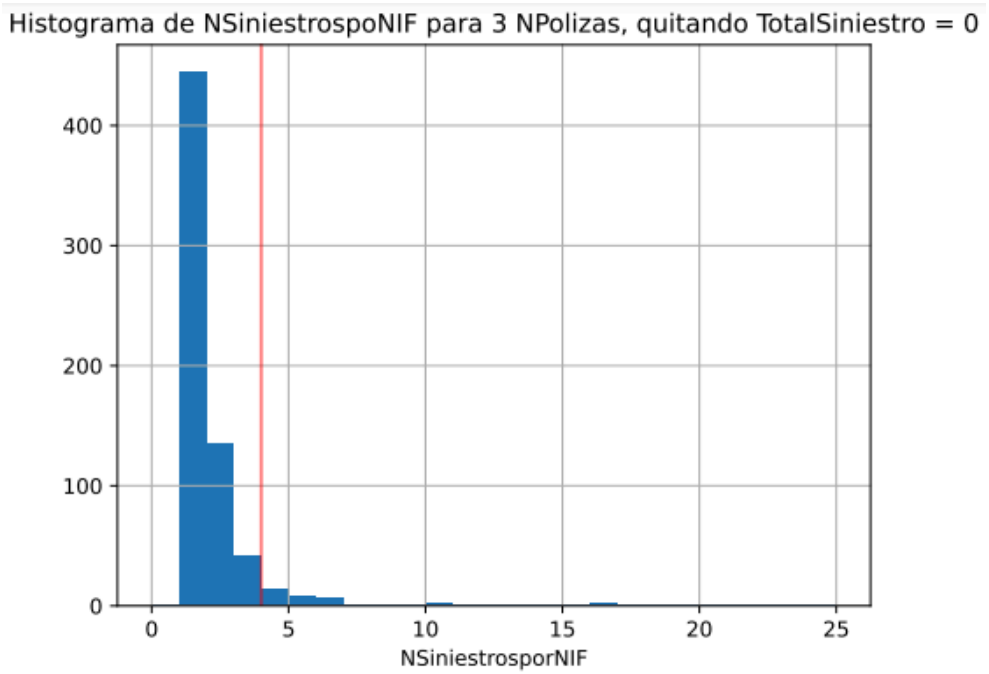
Este modelo se centra en estudiar cuánto tiempo pasa desde que se contrata una póliza de seguro hasta que se presenta un siniestro de seguro. Para hacer esto, se toma en cuenta el número de siniestros en relación con el número de pólizas para cada día.

Luego, se crea un gráfico para visualizar estos datos. En este gráfico, se espera ver un valor constante, excepto al principio. Si hay una pequeña curva por encima de este valor constante al principio, significa que ha habido más siniestros de lo normal en los primeros días después de contratar la póliza. Esto podría indicar un posible fraude, como, por ejemplo, alguien que contrata una póliza de seguro con la intención de reclamar un siniestro que ya existía.

Para facilitar la visualización de los datos, se agrupan en bloques de un año. Según el gráfico, la curva no se estabiliza hasta el quinto año. Sin embargo, el modelo devuelve un campo que indica si ha pasado un año o menos desde que se contrató la póliza hasta que se presentó el siniestro, ya que esto es lo que más interesa para detectar posibles fraudes.

Modelo 2: El modelo de número de pólizas

Este modelo organiza los datos según el número de pólizas de seguro que tiene cada cliente. Para cada número de pólizas, se agrupan los clientes y se calcula un valor llamado "percentil" basado en el número de siniestros que han hecho. Si el valor del percentil es igual al valor más alto posible, entonces no se considera un valor atípico (o "outlier") y no se marcará como fraude.

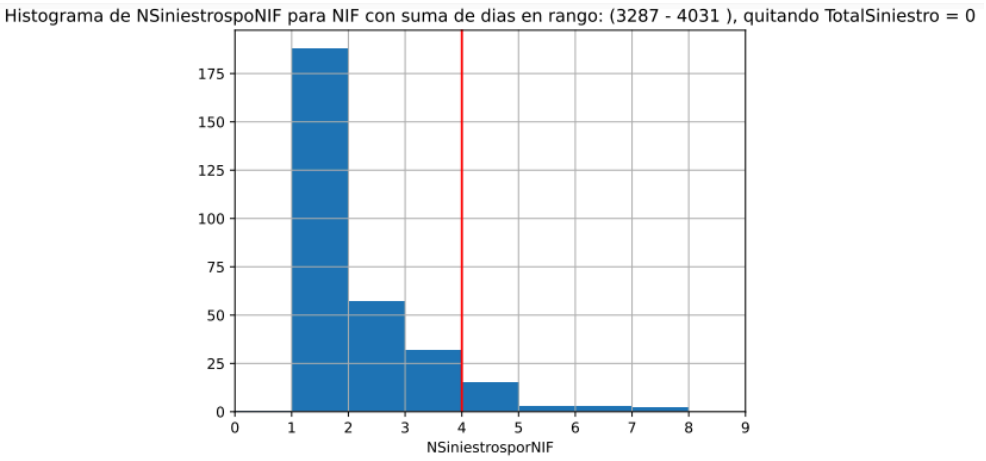


El modelo también puede crear un gráfico para ayudar a visualizar los datos. Sin embargo, hay que tener en cuenta que el número de clientes con un cierto número de pólizas y siniestros puede variar, lo que puede afectar cómo se ve el gráfico.

Por eso, el modelo también crea histogramas (que son un tipo de gráfico) para cada número de pólizas, siempre que haya suficientes clientes con ese número de pólizas. El número mínimo de clientes necesario para crear un histograma se puede ajustar en el archivo de configuración del modelo.

Modelo 3: Modelo de tiempo de pólizas

Este modelo organiza los datos según el tiempo total que las pólizas de seguro de cada cliente han estado en vigor. Para cada rango de tiempo, que se determina por un número de agrupación que se establece en el archivo de configuración del modelo, se calcula un valor llamado “percentil” basado en el número de siniestros que ha hecho cada cliente.

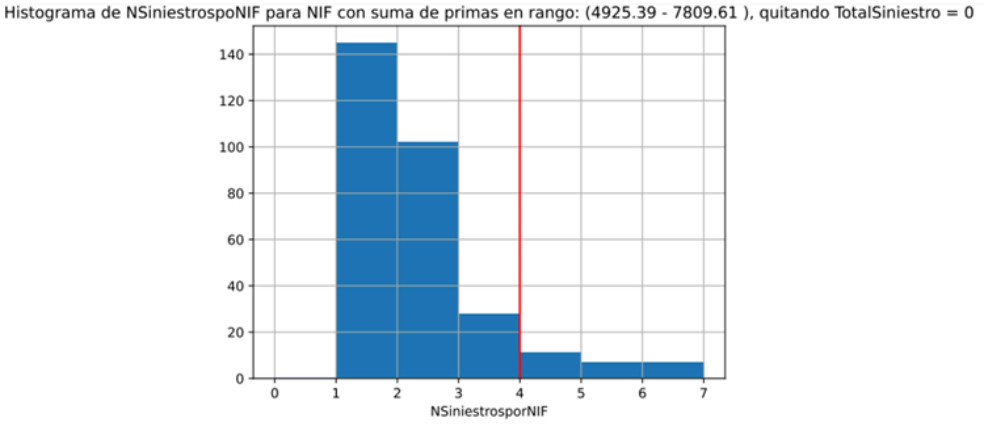


Si el valor del percentil de los clientes aumenta, el rango de tiempo será mayor. En el gráfico que se genera, se muestran puntos rojos que están por encima del percentil, y se consideran valores atípicos (o "outliers").

Modelo 4: Modelo de suma de primas

Este modelo organiza los datos según la suma total de las primas de seguro de cada cliente. Para cada rango, que se determina por un número de agrupación que se establece en el archivo de configuración del modelo, se calcula un valor llamado "percentil" basado en el número de siniestros que ha hecho cada cliente.

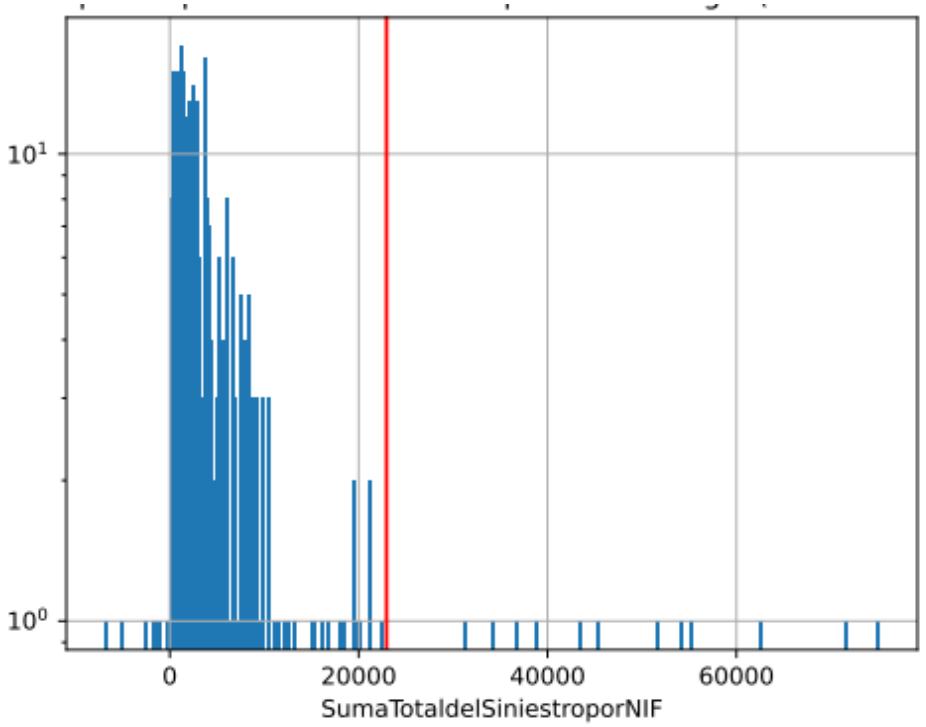
Si el valor del percentil de los clientes aumenta, el rango será mayor. En el gráfico que se genera, se muestran puntos rojos que están por encima del percentil, y se consideran valores atípicos (o "outliers").



Modelo 5: Modelo de suma de primas con valor del siniestro

Este modelo de seguro analiza la información de los clientes basándose en la suma total de las primas de sus pólizas. Para cada grupo de clientes con una suma similar de primas, se calcula un valor llamado "percentil" basado en la suma total de los siniestros que han hecho.

Si el valor del percentil aumenta, significa que el rango de suma de primas también será mayor. Los clientes que tienen más siniestros que el valor del percentil se consideran inusuales.



Resultados obtenidos

Modelo 1: El modelo de declaración de siniestro desde fecha de alta de la póliza

Este modelo procesa los años que han pasado desde el alta de la póliza hasta la declaración del siniestro. El cociente es el número de siniestros entre el número de pólizas.

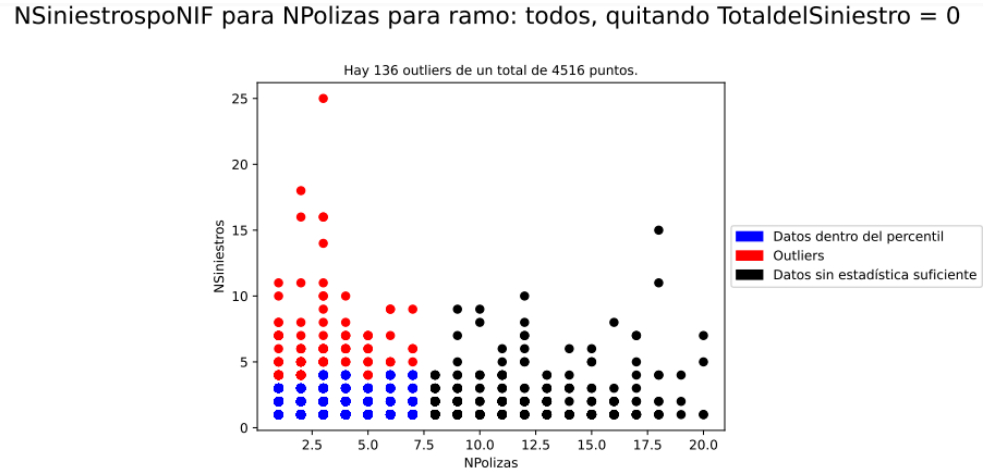
Los resultados muestran que bajo el Modelo 1, **el 7,16 % de los siniestros tendría una señal de alarma** de fraude de forma individual. En comparación con los demás modelos, es el que presenta el porcentaje más alto. Este es el único modelo que se presenta en combinación de los otros modelos, de forma que solamente presenta alarma si existe por lo menos un modelo más que ha presentado alarma también.

Modelo 2: El modelo de número de pólizas

Este modelo agrupa los datos por el número de pólizas que tiene cada cliente. Por cada número de pólizas, se agrupan los clientes y se calcula el percentil con el número de siniestros. En caso que el valor del percentil sea igual que el máximo valor, no se considerará como outlier y no se marcará como fraudulento. También se puede guardar una gráfica para una visualización de los datos.

El modelo 2 marca una señal de alarma **en un 4,26% de los siniestros hay un aviso de posible fraude**.

En este modelo observamos a los clientes cuyo número de pólizas frente a los siniestros está en rojo son los que se proporcionan en el fichero de salida.



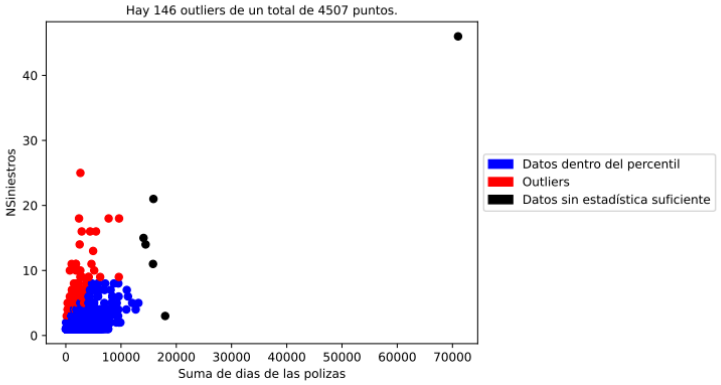
En los puntos azules están los clientes que estarían dentro de parámetros de no señal de fraude. Estos clientes corresponden a las frecuencias esperadas de siniestros. Los puntos en rojo son los clientes que se escapan a la frecuencia esperada, donde se da una señal de alarma. Los puntos negros corresponden a categorías de alto número de pólizas sobre los que no hay datos suficiente para establecer márgenes de frecuencia.

Modelo 3: Modelo de tiempo de pólizas.

Este modelo agrupa los datos por la suma del tiempo que han estado en vigor todas las pólizas de cada cliente. Por cada rango de tiempo, marcado por el número de agrupación puesto en el fichero de configuración, se calcula el percentil con el número de siniestros de cada cliente. Si el valor de los bloques de los clientes aumenta, el rango será mayor (por ejemplo, de 1 día a 1000 días).

El modelo 3 marca una señal de alarma **en un 3,82% de los siniestros hay un aviso de posible fraude**.

NSiniestrospoNIF para suma de días de las pólizas por NIF, quitando TotaldelSiniestro = 0



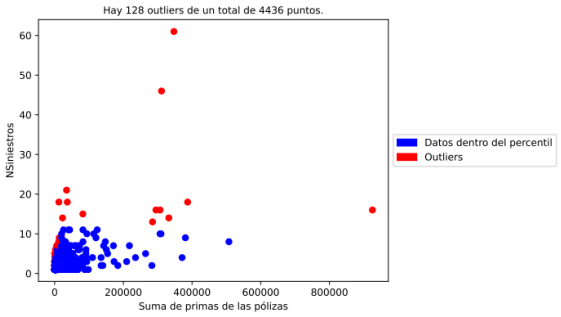
En este gráfico observamos las categorías de no fraude en color azul, y las categorías de fraude en color rojo. Estas categorías son combinaciones de número de siniestros, y antigüedad total de días de las pólizas sumadas. En las categorías de color negro no hay datos suficientes para establecer intervalos de frecuencia.

Modelo 4: Modelo de suma de primas

Este modelo agrupa los datos por la suma de las primas de las pólizas de cada cliente. Por cada rango, marcado por el número de agrupación puesto en el fichero de configuración, se calcula el percentil con el número de siniestros de cada cliente.

El modelo 4 marca una señal de alarma **en un 2,84% de los siniestros hay un aviso de posible fraude**.

NSiniestrospoNIF para suma de primas de las pólizas por NIF para ramo: todos, quitando TotaldelSiniestro = 0



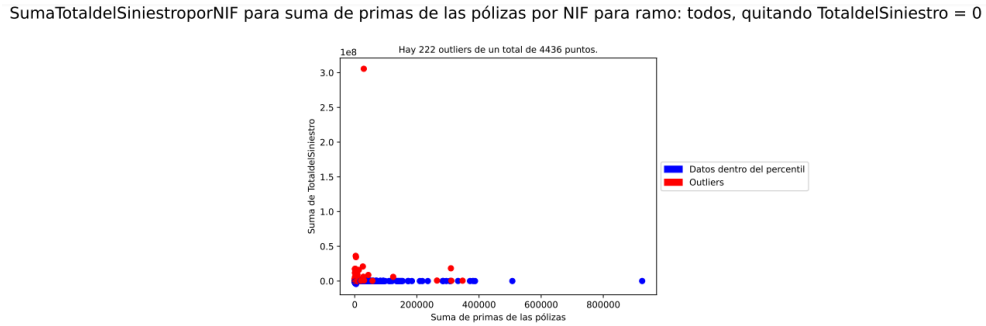
En este gráfico observamos, como en los anteriores modelos, las categorías de no fraude en color azul, y las categorías de fraude en color rojo. Estas categorías son combinaciones de

número de siniestros, y valor total de las primas por cliente. Vemos que este es uno de los modelos más exigentes, de ahí que se observan pocas categorías rojas.

Modelo 5 Modelo de suma de primas con valor del siniestro

Este modelo agrupa los datos por la suma de las primas de las pólizas de cada cliente. Por cada rango, marcado por el número de agrupación puesto en el fichero de configuración, se calcula el percentil con la suma de los siniestros producidos por cada cliente.

El modelo 5 marca una señal de alarma **en un 4,89% de los siniestros hay un aviso de posible fraude**.



En este gráfico observamos, como en los anteriores modelos, las categorías de no fraude en color azul, y las categorías de fraude en color rojo. Estas categorías son combinaciones de el valor total de los siniestros pagados, y valor total de las primas por cliente. Vemos que este es el único modelo que relaciona dos valores económicos, ya que todos los anteriores realizaban combinaciones con el número de siniestros.

Fichero de salida

Cada modelo cualifica al cliente como fraude/ no fraude, con valores booleanos. Sobre la consulta se extrae un fichero de salida de los modelos está en formato csv y muestra para cada cliente en qué modelos ha sido marcado como posible fraudulento para una posible investigación posterior.

El valor de exigencia de los modelos se pueden modificar, y en este caso el nivel de exigencia para considerar un outsider, es decir, un valor que está fuera de lo normal, está por encima del 95% de los casos.

Una sola señal de riesgo no es suficiente, pero más de 1 señal sí que es indicador más fehaciente de posible fraude.

La distribución de las combinaciones de marcas o señales de posible riesgo lo vemos en la siguiente tabla.

Nivel fraude	%	Acumulado
5	0,0%	0,01%
4	0,5%	0,55%
3	2,3%	2,82%
2	5,1%	7,90%
1	3,8%	11,69%
0	88,3%	100%

En este gráfico vemos que el nivel más bajo de calificación de fraude, es decir, que solo acumula una señal en los 5 modelos, se da en un 3,85% de **los siniestros**.

Lo más correcto es ser más exigente, y si accedemos a definir por lo menos tener dos señales de fraude para iniciar una revisión, obtendremos el 7,9% de los siniestros. Si elevamos a **3 señales para iniciar una inspección, obtenemos que un 2,82% son sospechosos de fraude**.