UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MÁSTER UNIVERSITARIO EN CIENCIAS ACTUARIALES Y FINANCIERAS

Trabajo Fin de Máster

Regresión Logística Con Datos Asimétricos: Aplicación a la detección de Fraude en el momento de la suscripción en seguros multi-riesgo (Hogar).

AUTOR: Johana Osorno Gómez
TUTOR: Zuleyka Díaz Martínez

CURSO ACADÉMICO: 2018/2019

DECLARACIÓN DE NO PLAGIO

Dña. **Johana Osorno Gómez** con **NIE Y4113355-A**, estudiante de Máster en la Facultad de Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid en el curso **2018-2019**, como autora del trabajo de fin de máster titulado **"Regresión Logística Con Datos Asimétricos: Aplicación a la detección de Fraude en el momento de la suscripción en seguros multi-riesgo (Hogar)",** y presentado para la obtención del título correspondiente, cuya tutora es **Zuleyka Díaz Martínez.**

DECLARO QUE:

El trabajo de fin de máster que presento está elaborado por mí y es original. No copio, ni utilizo ideas, formulaciones, citas integrales e ilustraciones de cualquier obra, artículo, memoria, o documento (en versión impresa o electrónica), sin mencionar de forma clara y estricta su origen, tanto en el cuerpo del texto como en la bibliografía. Así mismo declaro que los datos son veraces y que no he hecho uso de información no autorizada de cualquier fuente escrita de otra persona o de cualquier otra fuente.

De igual manera, soy plenamente consciente de que el hecho de no respetar estos extremos es objeto de sanciones universitarias y/o de otro orden

Firmado en Madrid, a **14 de septiembre de 2019.**

Fdo.: Johana Osorno Gómez

ÍNDICE

ÍNDICE DE IMÁGENES

ÍNDICE DE RESULTADOS

ÍNDICE DE TABLAS

1. RESUMEN

En este trabajo se construye un modelo de regresión logística para hallar la probabilidad de fraude de una póliza nueva, en el ramo de hogar, a través de los datos que se obtienen en el momento de la suscripción de la misma.

Los datos utilizados fueron proporcionados por una importante entidad aseguradora y corresponden a una muestra de 20 variables categóricas y 38.240 pólizas nuevas de los años 2017 y 2018, en donde el 4.7% ha sido identificado como pólizas con fraude (es decir, pólizas que han comunicado algún siniestro con intensión de fraude[1]). Es importante resaltar que este porcentaje de fraude sobre el total de la muestra puede considerarse alto si se tiene como referencia el 1.88%, que corresponde a la media del total de fraude (en todos los ramos), a nivel nacional (AXA España, 2019).

Antes de comenzar con el modelo, se realiza una leve introducción de las implicaciones del fraude en los seguros y se muestran algunas cifras correspondientes al año 2018. Esto con el fin de dar una idea de la importancia de este tipo de modelos en las compañías aseguradoras.

El desarrollo del modelo comienza con un análisis descriptivo de los datos y la evaluación de la relación de cada variable predictora con la variable respuesta (Fraude). Luego se realiza un muestreo estratificado para dividir el conjunto de datos en dos subconjuntos, uno para entrenamiento del modelo y el otro para su validación. Sobre el conjunto de datos de entrenamiento, se realizan dos pruebas para la selección de las variables significativas (Algoritmo Boruta (Miron B. & Witold R. , 2010) y Método de extracción AIC (R Core Team, 2019)) y después de determinar estas variables, se ejecuta el modelo GLM de regresión logística. Los resultados de este modelo se implementan, inicialmente, sobre los datos de entrenamiento para identificar si se requieren ajustes antes de evaluarlo en el conjunto de validación. Por ser un conjunto de datos asimétrico, no es conveniente determinar el nivel de probabilidad de pertenencia a la categoría Fraude como el 50% (como generalmente se hace), es por esto que se realiza una prueba de optimización de costos, utilizando un caso base como referencia, con la que se busca minimizar la pérdida monetaria (en la que incurriría la empresa por utilizar el modelo), en función de los errores y aciertos que comete el modelo en la clasificación. El resultado de esta optimización es el valor de la probabilidad de fraude con la que se minimizan los costos por mala clasificación del modelo, y a partir de ésta se construye la matriz de confusión con sus respectivas métricas, la curva ROC con el valor del AUC (área bajo la curva), y se

[1] En este trabajo no se distingue ninguna característica del tipo de fraude (si es un fraude total, parcial, garantía afectada, valor de reclamación de fraude, etc.)

calcula el efecto monetario (pérdidas/ganancias) sobre el caso base. Cuando el modelo se acepta como buena herramienta de clasificación, se procede a implementarlo en el conjunto de validación utilizando el mismo procedimiento (teniendo como referencia el valor de probabilidad determinado en el entrenamiento del modelo), esto es, se construye la matriz de confusión con sus métricas, la curva ROC con el cálculo del AUC y el efecto monetario (pérdidas/ganancias) sobre el caso base. Al final se comparan los resultados de ambos conjuntos para demostrar que es un buen modelo de clasificación.

Después de aceptar el modelo como buen clasificador, se procede a construir la función logit que servirá para la clasificación, en función del fraude, de las pólizas nuevas que vaya teniendo la compañía y se presentan 4 ejemplos de clasificación para explicar cómo funciona: 2 de pólizas que son fraude y 2 que no lo son.

Para finalizar, se propone otra forma de determinar el valor de la probabilidad que puede ser más fácil de entender e implementar. Esta prueba consiste en evaluar los histogramas de las distribuciones de probabilidad de cada categoría de la variable respuesta. El objetivo de esta prueba no es proponer un único valor de probabilidad, sino una serie de rangos en los que la decisión del valor de la probabilidad a tomar como referencia depende de los objetivos del análisis que se realice (recomendado principalmente para objetivos de análisis generales de clasificación y validación, no recomendado para acciones puntuales que impliquen inversión económica).

2. INTRODUCCIÓN

Es común pensar en un contexto negativo cuando se hace alusión a la palabra **fraude**, desde la definición general que realiza el diccionario del español jurídico de la real academia española de la lengua: "*Acción contraria a la verdad y a la rectitud que perjudica a la persona contra quien se comete - en determinadas circunstancias puede ser constitutiva de delito*" (DEJ, 2019), hasta la definición particular para el sector seguros que hacen en CEA en donde lo asemejan a una **estafa.** (Martínez de la Puente, s.f.)

En el sector asegurador, el fraude puede ser cometido por cualquiera de las partes que intervienen: tomador y/o beneficiario de la póliza, agentes de venta, reparadores, etc., y cuando se hace referencia al asegurado, todas las definiciones conllevan a entender el fraude como una actividad ilícita en la que éste pretende obtener un beneficio al que no tiene derecho, ya sea reportando daños de forma exagerada, cubriendo daños antiguos con un siniestro actual, solicitando coberturas que no se han contratado, etc. (Ayuso, 1998; Díaz Sanjuán, 2018; Sector Asegurador, s.f.)

Las personas que realizan fraude generalmente piensan que están aseguradas por compañías grandes y sólidas, y que sus reclamaciones adicionales, a las cuales no tienen derecho, no tendrán ningún efecto, sin embargo, en sectores como el asegurador donde los precios son calculados generalmente para un colectivo, cuando una persona perteneciente a ese colectivo realiza un fraude, lo que está haciendo realmente es enriqueciéndose con el dinero que pagan los asegurados honestos en las pólizas de ese colectivo, esto se debe a que cuando la compañía aseguradora realiza la asignación de precios (tarificación), traslada todos esos costos adicionales a sus clientes, quienes se ven obligados a pagar primas más altas por una misma cobertura. (Álvarez, 2018; AXA España, 2019; De la Espriella, 2012; Martínez de la Puente, s.f.; UNESPA, 2019)

En el ámbito español, la tasa de fraude es creciente, pasando de menos del 1% en 2012 a un 1,88% en 2018. La evolución anual se muestra en la siguiente gráfica, en la que se puede ver claramente que, en estos 6 años, la tasa de fraude se ha duplicado:

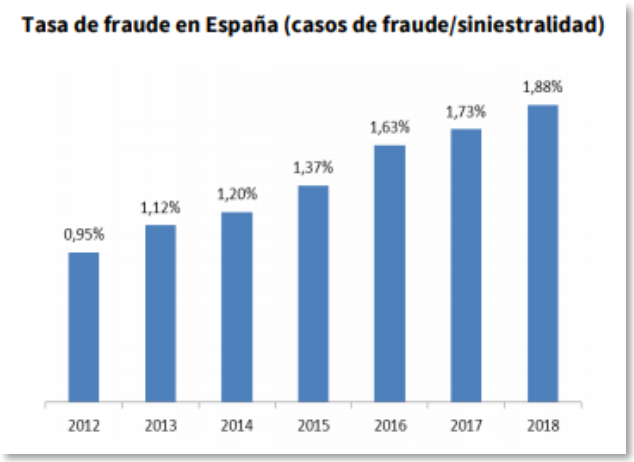


Imagen 1: Evolución de la tasa de fraude en España.
Fuente: (AXA España, 2019)

En cuanto al fraude por tipo de ramo, el de automóvil es el que presenta mayor porcentaje de fraude detectado (esto está completamente ligado a que es el ramo que más presencia tiene debido a la obligatoriedad que hay para los vehículos automotores de contar con un seguro para poder circular) (Soria, 2019), seguido de los ramos denominados "Multi-riesgo" en los que se incluyen los seguros de hogar, comercio, oficinas, comunidades y embarcaciones. Es de resaltar que, mientras para el año 2017 el ramo automóvil representaba alrededor de un 60% del total y los multi-riesgo alrededor del 30%, (AXA España, 2018), para el año 2018, automóvil ha disminuido su porcentaje a un 49% mientras que los multi-riesgo han incrementado su porcentaje hasta un 40%, incremento impulsado particularmente por fraudes en el ramo de hogar (AXA España, 2019), donde los fraudes más frecuentes se dan en las reclamaciones por averías mecánicas de electrodomésticos y falsos robos. (Álvarez, 2018; Equipo de redacción MV,s.f.; Grimaldi, 2019). El otro 11% restante corresponde al resto de ramos diversos. De forma gráfica sería:

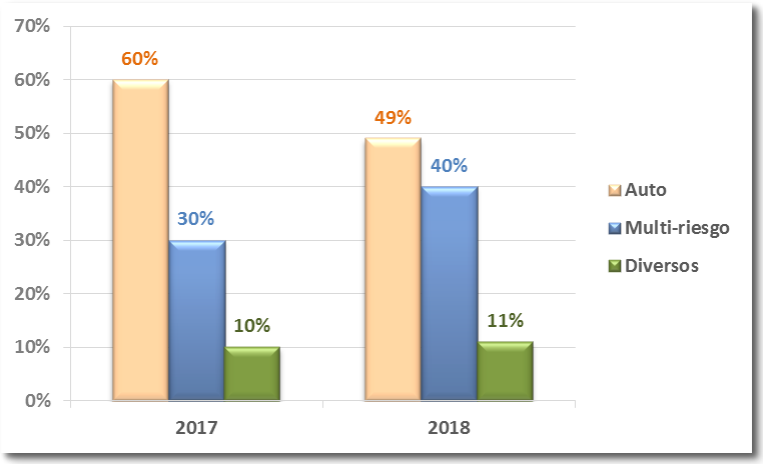


Imagen 2:Distribución del fraude por Ramos
[Elaboración propia. Fuente: (AXA España, 2018; AXA España, 2019)]

Esta distribución del fraude en los diferentes ramos es la que ha impulsado a que la mayor parte de los estudios del fraude estén enfocados en el ramo de automóvil, principalmente estudios de predicción de fraude a partir de la ocurrencia de un siniestro. (Ayuso, Guillen, & Artís, 1999; Díaz Sanjuán, 2018; García, 2017). Son pocos los estudios que se han hecho de los ramos multi-riesgo e incluso menos los que pretenden explicar o predecir el fraude desde el momento de la suscripción[2]. Ésta fue la principal motivación que impulsó el desarrollo de este trabajo.

Ahora, para mostrar de forma general el problema que se aborda en este análisis, es necesario tener en cuenta que, en el sector asegurador, una de las grandes características de los datos es que son bastante asimétricos (o también conocidos como datos desbalanceados). Esta característica se presenta debido a la gran aleatoriedad que existe a la hora de la materialización de un riesgo y se convierte en todo un reto para el análisis de los datos y modelamiento de los mismos en cualquier parte del proceso que tiene una póliza en su duración. (De la Espriella, 2012).

Cuando se realizan modelos de predicción sobre datos asimétricos, la conclusión de si el modelo obtenido es bueno o no, no puede soportarse sobre las mismas medidas (baremo) utilizadas en el análisis de conjuntos de datos simétricos, tal es el caso del criterio AUC (área bajo la curva ROC[3]) , que siempre se busca que sea alto y en datos desbalanceados es normal que sea bajo (Charpentier, 2019) , o en las matrices de confusión en donde la medida de exactitud es suficiente para concluir si la capacidad predictiva del modelo es buena (alta) o no, sin embargo, en datos desbalanceados, no es suficiente ni concluyente obtener un valor alto de exactitud sin revisar el resultado e impacto de los demás resultados que se obtienen de la matriz. (Brownlee, 2015)

En este trabajo se persigue un objetivo fundamental: determinar el valor de la probabilidad a partir del cual se considerará que una póliza puede ser clasificada como posible futuro fraude a partir de los datos que se obtienen al momento de la suscripción. Para lograr este objetivo, este trabajo se desarrolla bajo el siguiente orden: en el primer apartado se realiza un resumen de lo que se desarrolla en el trabajo. En el segundo, se expone la introducción con datos generales del seguro en España, los diferentes ramos y orden seguido en la modelación. A partir del tercer apartado se comienza con la explicación del modelamiento exponiendo primero la teoría general de los GLM y, en particular, de la regresión logística y las medidas de bondad de ajuste. Luego, en el cuarto apartado, se comienza con el análisis descriptivo de los datos en el que se muestran las distribuciones de cada una de las

[2] Lo más probable que es existan este tipo de análisis, sin embargo, no fue posible encontrarlos cuando se realizó la revisión bibliográfica.
[3] Las siglas ROC corresponden a Receiver Operating Characteristic Curve.

categorías de cada variable y se realiza el test Chi-cuadrado para validar la hipótesis de independencia entre cada variable y la variable respuesta: Fraude. En el quinto apartado, se explica cómo se realiza la división del conjunto de datos en los subconjuntos de entrenamiento y validación. En el sexto, se realizan dos pruebas para la selección de las variables que realmente son significativas en el modelo de predicción, la primera prueba consiste en la ejecución del algoritmo Boruta (se basa en el algoritmo *random forest* realizando una comparación entre las variables originales y unas variables auxiliares creadas por el mismo algoritmo), y la segunda es la ejecución del método de extracción AIC, que se genera con el propio GLM (como su nombre lo indica, se basa en la optimización del AIC), ambas pruebas pertenecientes a los métodos envolventes que utilizan procedimientos paso a paso (Guerrero, 2016). Después de seleccionar el modelo óptimo (modelo GLM con las variables significativas), en el apartado 7, se procede a la obtención de los resultados del modelo con el conjunto de entrenamiento y se realizan las predicciones. Para proponer el valor de la probabilidad a través de la cual se determina si una póliza es o no Fraude, se realiza un análisis de optimización de costos en función de un caso base propuesto, en el que se asignan valores monetarios a los errores que comete el modelo, y se busca la menor pérdida posible, la cual está asociada a un valor de probabilidad que será el propuesto como punto de corte para la clasificación. Después de hallar este valor de probabilidad, se procede con la construcción de la matriz de confusión y las métricas asociadas a ésta, la curva ROC, el valor AUC, y se calcula el efecto monetario por implementación del modelo. En el apartado 8 se realiza la evaluación del modelo obtenido con el conjunto de entrenamiento aplicándolo al conjunto de validación, para lo cual se realizan las predicciones de probabilidad a cada póliza y se calcula la matriz de confusión y las métricas asociadas a ésta, la curva ROC, el valor AUC, y se calcula el efecto monetario por implementación del modelo en este conjunto de datos. Para finalizar esta sección, se realiza una comparación de los resultados obtenidos con el modelo en ambos conjuntos, el de entrenamiento y el de validación, esto con el fin de concluir si el modelo obtenido es un buen modelo de clasificación de fraude o no lo es. En el apartado 9 se define la función logit obtenida por el modelo seleccionado y se presentan cuatro ejemplos de clasificación, dos para pólizas que realmente son fraude y dos para pólizas que no lo son. En el apartado 10, se propone una forma gráfica para determinar el valor de la probabilidad, la cual puede ser más sencilla que realizar la optimización de costos. Con esta propuesta gráfica no se da un único valor de probabilidad, sino que se proponen rangos para que la selección de este valor sea a discreción de la compañía que los usa de acuerdo a los objetivos que tenga con el uso de esta herramienta.

Para finalizar el trabajo se presentan las conclusiones, posibles futuros trabajos, bibliografía y anexos, en estos últimos se incluye el código utilizado en el software R (RStudio Team, 2018), para la modelación.

3. MODELOS GLM

Como se mencionó anteriormente, una de las grandes características de los datos en el sector asegurador es que son bastante asimétricos, esto es: el número de observaciones no es homogéneo para todas las clases de un conjunto de datos y por lo general hay una variable mayoritaria (con la mayor proporción del total de datos) (Santacruz, 2016).

A la hora de modelar este tipo de datos es importante conocer bien los modelos que se pretenden usar y las características propias de cada uno de ellos, un ejemplo son los modelos de *Random Forest de Machine Learning* los cuales son sensibles ante las proporciones de las diferentes clases, por lo que este tipo de algoritmos tienden a favorecer la clase con la mayor proporción de observaciones y producir resultados sesgados (Santacruz, 2016), es así como la mayor exactitud se obtiene en los valores correspondientes a esa variable mayoritaria. Esto se convierte en problema cuando el objetivo es entender y tratar de explicar el comportamiento específico de la variable minoritaria en vez de tratar de minimizar la tasa de error global o maximizar la capacidad predictiva.

Un mecanismo muy utilizado para el procesamiento y análisis de datos asimétricos son los denominados GLM (Generalized Linear Models), estos modelos son una extensión de los modelos lineales clásicos y se caracterizan porque permiten modelar, como variables respuesta, aquellas que tienen modelos de distribución de errores distintos a una distribución normal. *"El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho*" (Wikipedia, s.f.) .

La generalización del modelo lineal la realiza en varias direcciones (Díaz, 2019):

- La variable respuesta (Y) sigue una distribución de probabilidad de la familia exponencial, en donde la distribución normal es considerada como un caso particular.
- La esperanza de la variable respuesta ya no es directamente el **predictor lineal**, sino que está relacionada con él a través de la **función de enlace** (conocida como "*link function*")
- La varianza de la variable respuesta no necesariamente es constante, sino que es función de su media y se denomina **función varianza.**

3.1. Modelos de regresión logística[4]

Estos modelos GLM tienen la particularidad de que su variable respuesta es medida a través de una escala binaria, generalmente medida como "éxito" o "fracaso". La forma general en la que se define esta variable binaria es:

$$Z = \begin{cases} 1, si\ la\ salida\ es\ un\ éxito \\ 0, si\ la\ salida\ es\ un\ fracaso \end{cases}$$

Con probabilidades $P(Z = 1) = \pi$, y $P(Z = 0) = 1 - \pi$.

Para el caso particular en el que las $\pi_j's$ son iguales, se define una nueva variable Y tal que:

$$Y = \sum_{j=1}^{n} Z_j$$

Donde Y representa el número de "éxitos" en n "ensayos", por lo que la distribución de la variable aleatoria Y se corresponde con una $binomial\ (n, \pi)$.

Para generalizar este modelo a un número N de variables aleatorias independientes, $(Y_1, Y_2, \dots, Y_N$, correspondientes al número de sucesos en N diferentes subgrupos o clasificaciones), se describe la proporción de éxitos como $P_i = \frac{Y_i}{n_i}$, en cada subgrupo en términos de los niveles de factor o variables predictoras, donde $E(Y_i) = n_i\pi_i$, o lo que es lo mismo (pero en términos de probabilidad), $E(P_i) = \pi_i$, obteniendo como modelo de probabilidades:

$$g(\pi_i) = x_i^T \beta$$

Donde x_i es el vector de variables predictoras (variables categóricas con diferentes niveles de clasificación), β: el vector de parámetros (coeficientes de las variables x_i) y g: función de enlace.[5]

Para este trabajo, en donde el objetivo principal es modelar la variable Fraude en función de una serie de variables predictoras de tipo categórico, se utilizará un modelo logit, con función de enlace η , para realizar dicha clasificación y obtener el valor de la probabilidad sobre la cual una póliza pueda ser objeto de investigación por posible futuro fraude. Esta función logit se representa como:

[4] Para dar esta definición se utilizan las fórmulas y definiciones expuestas por (Dobson, 2001; Morales, 2018)
[5] El caso básico de este tipo de modelos es la regresión lineal simple: $\pi = x^T\beta$

$$\eta = \log\left(\frac{\pi}{1-\pi}\right)$$

El cual, en términos del modelo de probabilidades, es tal que:

$$\eta = \ logit\ (\pi_i) = \log\left(\frac{\pi}{1-\pi}\right) = \ x_i^T \beta$$

Siendo π la probabilidad de que el individuo tome el valor de 1 en la variable dicotómica (Fraude).

Al igual que en los modelos de regresión tradicionales, los resultados del modelo se basan en la interpretación del $valor\ p$ asociado y de los coeficientes de las variables predictoras:

- En cuanto a **la interpretación del $\boldsymbol{valor\ p}$, ésta es similar** a la del modelo lineal tradicional en donde se compara el resultado con un nivel de significatividad $\alpha = 0.05$, tal que, si $valor_p > \alpha$ se concluye que la variable no es significativa.

- En cuanto a **la interpretación de los coeficientes, ésta sí cambia** en este tipo de modelos ya que el modelo GLM no ajusta la variable respuesta sino **la función de enlace** (η) descrita anteriormente.
  Al cociente $\left(\frac{\pi}{1-\pi}\right)$ de la función de enlace, se le conoce como **odds ratio**, por lo que los coeficientes del modelo logit se deben interpretar como el **logaritmo del odds ratio**.

Una forma de facilitar la interpretación de los coeficientes es aplicándole la función inversa al modelo GLM (en este caso la función exponencial), de tal forma que se tiene una explicación del riesgo en un modelo multiplicativo donde cada factor de riesgo tiene una ponderación en el riesgo total, esto es:

$$odds = \ e^{\beta_0} * \ e^{\beta_1 X_1} * \ldots * e^{\beta_i X_i}$$

Generalmente, interesa hallar la probabilidad de pertenencia a la variable que se estudia (probabilidad de que un individuo, o caso, tome valor 1 en la variable respuesta), esto se realiza utilizando:

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

Donde

- π: hace referencia a la función de probabilidad
- η: combinación lineal de tipo $y = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$. Con "X" indicando el número de variables predictoras e "i" cada una de las pólizas consideradas en el análisis.

Es importante tener en cuenta que, por ser una función de probabilidad, π está acotada en el intervalo [0,1]. (Cañadas Reche, 2013; Díaz Sanjuán, 2018)

3.2. Medidas de bondad de ajuste

En cuanto a las medidas de bondad del ajuste, se debe tener en cuenta que para los casos particulares del GLM, modelos Logit y Probit[6], no tiene sentido utilizar el criterio del R^2 ya que por lo general este criterio es una medida que expresa el ajuste del modelo y en la regresión logística lo que se busca es una buena clasificación.

Los criterios más utilizados para medir este tipo de modelos de clasificación son (Heras, 2019):

- **Deviance**: la forma matemática de calcularla es

$$\Delta = 2(L_{max} - L(b_{MLE}))$$

  Donde L_{max}: máxima verosimilitud del modelo saturado, y $L(b_{MLE})$: verosimilitud del modelo que se está ajustando.

  El análisis de la deviance es una generalización del análisis de la varianza que se realiza en los modelos GLM y generaliza la suma de los cuadrados de los residuos de un modelo OLS. Esta medida es siempre positiva e indica la diferencia en el ajuste del modelo realizado con el del modelo que ajusta perfectamente los datos (conocido como modelo saturado). Cuando se comparan dos modelos a través de este criterio, **será mejor modelo el que tenga Menor Deviance**.

- **AIC (Criterio de Akaike):** la forma matemática de calcularlo es

[6] Para ver detalles del probit, consultar (Dobson, 2001; Heras, 2019; Díaz, 2019)

$$-2L(b_{MLE}) + 2k$$

Donde $L(b_{MLE})$: verosimilitud del modelo que se está ajustando, y k: número de variables regresoras.

Esta medida ayuda a evaluar el sobreajuste del modelo. Al incluir más variables regresoras en el modelo que se está ajustando, la verosimilitud del modelo sigue disminuyendo, sin embargo, como k incrementa, se evita el sobreajuste.
Por si sola esta medida no ofrece información sobre la calidad de un modelo, es un criterio que se utiliza para comparar modelos y concluir sobre cuál es el mejor, es decir, se utiliza como una medida de calidad relativa de un modelo estadístico para un conjunto de datos. Cuando se comparan dos modelos a través de este criterio, **será mejor modelo el que tenga Menor AIC.**

- **Chi-Cuadrado:** Este test es utilizado para analizar variables nominales o cualitativas. El resultado que se observa es el *valor p*, y se utiliza para determinar si se puede o no rechazar la hipótesis nula, la cual se identifica como *H_0: no existe ninguna asociación entre las variables*. Cuando el *valor p* ≤ 0,05 se rechaza H_0 y se concluye que las variables tienen una asociación estadísticamente significativa. Su expresión matemática es la siguiente:

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

- **Pseudo-R^2:** mide el grado de mejora en el ajuste del modelo del log de la verosimilitud respecto al modelo generalizado, es decir, sin variables predictoras. Generalmente este es la medida que se utiliza para establecer la capacidad predictiva de los modelos GLM, sin embargo, no es la más recomendada cuando se trata de un modelo en el que las variables predictoras son categóricas. Su expresión matemática es:

$$\frac{L(b_{MLE}) - L_0}{L_{max} - L_0}$$

3.2.1. Matriz de confusión

Otra medida de bondad de ajuste muy utilizada es la **Matriz de confusión**, la cual se utiliza en los modelos de aprendizaje automático supervisado para validar si el algoritmo está haciendo una asignación correcta de los casos en la clase que les corresponde o está confundiendo una clase por otra (Díaz Sanjuán, 2018). A partir de esta matriz se construyen diferentes criterios de medición que deben contemplarse como un conjunto y no por separado para evitar dar conclusiones sesgadas o realizar medidas que no son objetivo del estudio.

La estructura de la matriz de confusión es la siguiente:

	Estimado Positivo	Estimado Negativo	Total Real	
Real positivo	Verdadero Positivo *(VP)*	Falso Positivo *(FP)*		Error tipo I
Real negativo	Falso Negativo *(FN)*	Verdadero Negativo *(VN)*		
Total Estimación			Total Datos	
	Error Tipo II			

Imagen 3: Estructura de la matriz de confusión
[*Elaboración* propia]

Donde**:**

- **Estimado Positivo:** Lo que el modelo clasifica como Fraude.
- **Estimado Negativo:** Lo que el modelo clasifica como No Fraude.
- **VP:** Casos que el modelo clasifica como Fraude y en realidad son Fraude.
- **VN:** Casos que el modelo clasifica como No Fraude y en realidad son No Fraude.
- **FP:** Estos casos también son conocidos como **Error Tipo I** e identifica aquellos casos que el modelo clasifica como Fraude y no lo son. En aplicaciones como la realizada en este trabajo, el tener casos clasificados en FP implica costos económicos adicionales por incurrir en investigaciones que no resultan en fraude.
- **FN:** Estos casos también son conocidos como **Error Tipo II** e identifica aquellos casos que el modelo clasifica como No Fraude y terminan siendo Fraude. En aplicaciones como la realizada en este trabajo, el tener casos clasificados en FN implica costos económicos por fraudes realizados y no detectados (es decir, costos por falta de detección temprana).

A partir de esta matriz se generan los siguientes resultados, los cuales sirven para evaluar el desempeño del modelo:

Métrica	Definición	Forma de cálculo
Exactitud	También conocido como **Accuracy** e indica la exactitud global con la que predice el modelo. El inconveniente con esta medida es que si una categoría es sobrestimada y otra subestimada, el resultado puede dar un valor alto, que se interpretaría como bueno pero, por la estructura de los datos, no necesariamente ser así en función de la predicción de la categoría que interesa en el análisis.	$\frac{(VP+VN)}{Total\ Datos}$
Tasa de error	Indica la proporción de datos que han sido mal clasificados.	$\frac{(FP+FN)}{Total\ Datos}$
Sensibilidad	Indica el porcentaje que el modelo logra explicar de la categoría Fraude.	$\frac{VP}{Total\ Reales\ Positivos}$
Especificidad	Indica el porcentaje que el modelo logra explicar de la categoría No Fraude.	$\frac{VN}{Total\ Reales\ Negativos}$
Precisión	Indica del total de los predichos como Fraude, cuántos realmente lo son.	$\frac{VP}{Total\ Estimados\ Positivos}$
Valor de predicción negativo	Indica del total de los predichos como No Fraude, cuantos realmente no lo son.	$\frac{VN}{Total\ Estimados\ Negativos}$

Tabla 1:Métricas generadas a partir de la matriz de confusión.
Elaboración propia. Fuente (Díaz Sanjuán, 2018)

Para la construcción de esta matriz, hay que tener en cuenta que, en un modelo con datos simétricos, la asignación de la probabilidad de pertenecer a una categoría u otra es del 50%, con esto lo que se pretende es dar la misma probabilidad al evento de ocurrencia (o presencia) que se evalúa en la variable respuesta, y al de no ocurrencia (no presencia). En este caso, como los datos son asimétricos, asignar una probabilidad del 50% no es adecuada y debe ajustarse a algún criterio que sea coherente con la información que se está analizando.

4. DESCRIPCIÓN DE LA BASE DE DATOS

4.1. Objetivo del análisis

Utilizar un modelo de regresión logística para hallar el valor de la probabilidad, con base en los datos que se obtienen al momento de la suscripción, a partir del cual, una póliza nueva en el ramo de hogar, puede ser clasificada como una póliza sobre la que se realizará un futuro fraude.

4.2. Naturaleza de los datos

La información para este trabajo fue proporcionada por una importante entidad aseguradora, con más de 100 años de experiencia en el sector y reconocida por varios años como primera marca mundial de seguros según el ranking de Interbrand[7]. La base de datos consiste en una **muestra del total** de pólizas aseguradas, para el ramo de hogar, en 2017 y 2018 (38.240 pólizas para el análisis), en donde 1.798 (4,7%) han sido identificadas como pólizas que han reportado algún tipo de siniestro con intención de fraude.

4.3. Descripción de las variables

Todas las variables utilizadas en el análisis son categóricas, en total 20 contando la variable respuesta, donde las categorías pueden representar clases específicas o rangos.

Las variables utilizadas en el análisis son las siguientes:

- **Cod_npol**: código asignado a cada póliza como número de identificación de la misma (número aleatorio asignado para el análisis).

- **Antigüedad de la vivienda:** número de años (en rangos) que han pasado desde la construcción de la vivienda hasta el día de la suscripción de la póliza.

- **Contenido**: indica el valor monetario (en rangos) por el que se ha asegurado el contenido de la vivienda.

[7] Valora las mejores marcas a partir de 3 criterios: desempeño financiero de los productos y servicios, influencia de la marca para forzar la elección del cliente y capacidad para imponer un precio o asegurar ganancias (Interbrand, 2018)

- **Continente:** indica el valor monetario (en rangos) por el que se ha asegurado el continente de la vivienda.

- **Forma de pago**: indica la forma en que el tomador de la póliza ha decidido realizar el pago de la prima, puede ser anual, trimestral, mensual o de forma irregular.

- **Nota bureau:** Probabilidad de impago de la prima agrupada en categorías de nivel: probabilidad de impago leve, moderada, importante, crítico, muy crítico, genérico, sin información.

- **Nota global:** es una clasificación interna que tiene la compañía en la que evalúa la probabilidad de que el tomador de la póliza pase de una categoría de nota bureau a una de mayor riesgo.

- **Clasificación del mediador:** clasificación interna que realiza la compañía a sus agentes. La categoría está en función del número de pólizas que venden e ingresos que generan.

- **Provincia**: lugar geográfico donde se encuentra ubicada la vivienda.

- **Superficie:** metros cuadrados que tiene la vivienda (indicada en rangos).

- **Territorial**: zona geográfica donde opera el mediador que realizó la contratación de la póliza.

- **Tipo de vivienda:** indica si la vivienda corresponde a un piso, chalet, casa tradicional u otro tipo.

- **Ubicación:** indica si la vivienda está en el casco urbano, urbanización o un despoblado

- **Tipo de mediador:** identifica si quien realizó la contratación de la póliza es un agente comercial, un empleado de la compañía (no comercial), o si la contratación se realizó sin intervención o contacto directo (Directo).

- **Tipo de usuario**: indica si el tomador del seguro es inquilino o propietario.

- **Tipo de uso**: indica si la vivienda es habitual, secundaria u otro tipo de uso.

- **Rehabilitada**: indica si la vivienda ha sido rehabilitada o no.

- **Edad del tomador**: años (en rangos) que tiene el tomador al momento de la suscripción de la póliza.

- **Hipoteca**: variable dicotómica que indica si la vivienda tiene o no una hipoteca al momento de la suscripción.

- **Valor del cliente:** corresponde a la clasificación interna que realiza la compañía de todos sus clientes en función de la rentabilidad que generan por todos los productos activos que tengan. Aunque este trabajo se basa en un análisis de pólizas nuevas en el ramo de hogar, se tiene que tener en cuenta que, si la persona que suscribe la póliza ya tiene contratado otro producto anteriormente, ya va a tener una clasificación en alguna de las categorías de esta variable, mientras que, si es el primer producto que adquiere con la compañía, este valor no será posible medirlo (asignarlo) hasta un tiempo determinado.

- **Fraude**: variable respuesta que se quiere estimar. Esta variable dicotómica indica si la póliza ha sido identificada como póliza con intención de fraude al reportar algún siniestro.

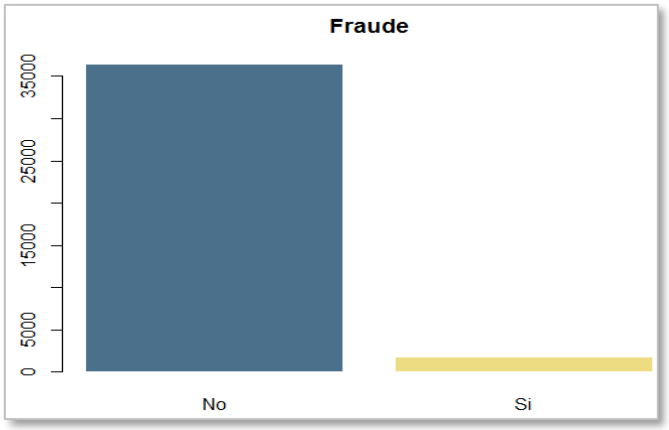
4.4. Análisis descriptivo de los datos

Antes de comenzar con el modelamiento de los datos es necesario realizar un análisis descriptivo de las variables para conocer un poco el comportamiento de las mismas y su relación con la variable de interés.

4.4.1. Variable de interés: Fraude

Para iniciar, se presenta la distribución de la variable respuesta: **Fraude**. Ésta es una variable dicotómica que se divide en dos categorías: Si o No (1 y 0 respectivamente). Cuando una póliza está clasificada dentro de esta variable como Si, es porque ha sido una póliza sobre la cual se ha reclamado un siniestro que ha sido identificado como fraude (independiente del tipo de fraude).

La distribución es la siguiente:



	Fraude	
	Cantidad	Proporción
No	36.442	95,3%
Si	1.798	4,7%
Total	38.240	100%

Imagen 4: Distribución de la variable respuesta: Fraude.
[Elaboración propia]

Como se puede observar en los resultados, son datos bastante asimétricos donde la categoría mayoritaria es No Fraude con un 95.3%.

Considerando que los datos pertenecen al ramo de hogar en donde los siniestros no son tan frecuentes, como sí puede suceder en un ramo como automóvil, se puede considerar como una muestra de fraude significativa para continuar con el análisis (en términos de proporción del total de datos).

Antes de iniciar el análisis descriptivo de las variables predictoras, es necesario resaltar que todas ellas son categóricas por lo que, además de mostrar sus frecuencias y porcentajes de fraude y no fraude dentro de cada una de las categorías, se debe realizar el test Chi-cuadrado para mirar la relación que existe entre cada una de ellas y la variable respuesta.

4.4.2. Antigüedad de la vivienda

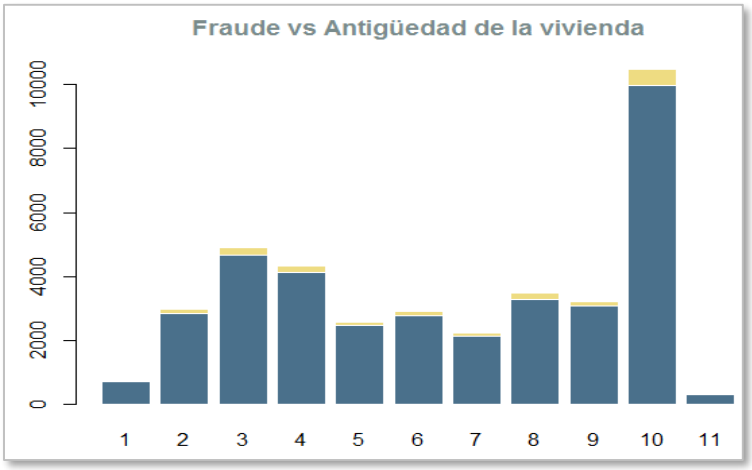


Imagen 5: Distribución variable antigüedad de la vivienda.
[Elaboración propia]

Tabla 2: Distribución variable antigüedad de la vivienda.

Antigüedad		Fraude				
Categoría	Rango	No	Si	Total	Prop. No	Prop. Si
1	0_ 5 años	735	30	765	96,1%	3,9%
2	5_ 10 años	2.839	136	2.975	95,4%	4,6%
3	10_ 15 años	4.680	216	4.896	95,6%	4,4%
4	15_ 20 años	4.116	226	4.342	94,8%	5,2%
5	20_ 25 años	2.467	121	2.588	95,3%	4,7%
6	25_ 30 años	2.794	129	2.923	95,6%	4,4%
7	30_ 35 años	2.138	108	2.246	95,2%	4,8%
8	35_ 40 años	3.293	189	3.482	94,6%	5,4%
9	40_ 45 años	3.071	147	3.218	95,4%	4,6%
10	>45	9.975	490	10.465	95,3%	4,7%
11	Sin Info	334	6	340	98,2%	1,8%
Total		36.442	1.798	38.240	95%	5%

En cuanto a la variable propiamente, se puede observar que la mayor proporción de viviendas que se aseguran (27.37%) tienen más de 45 años de construcción en el momento en el que se contrata la póliza, seguidos por las que tienen entre 10 y 15 años de construcción (12.8%). Además, es una variable que está bien poblada, es decir, el porcentaje de pólizas que no tienen este dato es bastante bajo (0.89%), lo que ayuda a que las estimaciones en las que se tenga en cuenta esta variable generen resultados que se pueden interpretar y explicar.

En cuanto a la presencia de fraude dentro de las categorías se puede ver que es muy homogénea y se encuentra alrededor del 4.4%.

Al realizar el test Chi-cuadrado para validar si existe o no asociación entre las variables, se obtiene lo siguiente:

Pearson's Chi-squared test		
data: antiguedad_vivienda and Fraude		
X-squared = 15.909	df = 10	p-value = 0.1023

Resultado 1:Test Chi-Cuadrado: Fraude vs. Antigüedad de la vivienda

Como el $valor\_p$ = 0.1023 es mayor al nivel de significatividad del 0.05, no se puede rechazar H_0, por lo tanto no se puede concluir que las variables antigüedad de la vivienda y fraude estén asociadas ya que no hay suficiente evidencia para concluirlo.

4.4.3. Contenido

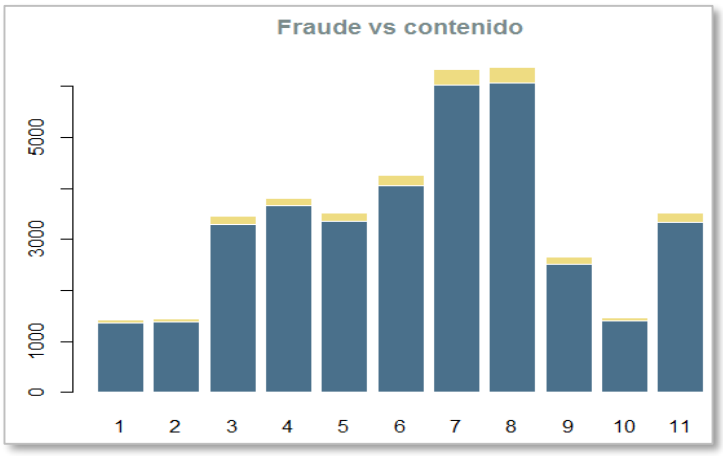


Imagen 6:Distribución variable contenido.
[Elaboración propia]

Contenido		Fraude				
Categoría	Rango	No	Si	Total	Prop. No	Prop. Si
1	0	1.365	52	1.417	96,3%	3,7%
2	>0_5000	1.390	58	1.448	96,0%	4,0%
3	5000_10000	3.297	154	3.451	95,5%	4,5%
4	10000_15000	3.665	148	3.813	96,1%	3,9%
5	15000_20000	3.347	164	3.511	95,3%	4,7%
6	20000_25000	4.047	209	4.256	95,1%	4,9%
7	25000_30000	6.014	313	6.327	95,1%	4,9%
8	30000_35000	6.062	304	6.366	95,2%	4,8%
9	35000_40000	2.520	134	2.654	95,0%	5,0%
10	40000_45000	1.406	63	1.469	95,7%	4,3%
11	>45000	3.329	199	3.528	94,4%	5,6%
Total		36.442	1.798	38.240	95%	5%

Tabla 3: Distribución variable Contenido

En cuanto a la variable Contenido se puede observar que la mayor proporción, en cuanto a rangos de contenido asegurado, se encuentra entre 25000 y 35000, categorías 7 y 8 con proporciones sobre el total de 16.55% y 16.65% respectivamente.

También se observa que es una variable que está bien poblada ya que ni siquiera existe una categoría que se defina como "sin información", lo que ayuda a que las estimaciones en las que se tenga en cuenta esta variable generen resultados que se pueden interpretar y explicar.

En cuanto a la presencia de fraude dentro de las categorías se puede ver que es muy homogénea y se encuentra alrededor del 4.6%.

Al realizar el test Chi-cuadrado para validar si existe o no asociación entre las variables, se obtiene lo siguiente:

Pearson's Chi-squared test		
data: contenido and Fraude		
X-squared = 20.664	df = 10	p-value = 0.02356

Resultado 2:Test Chi-Cuadrado: Fraude vs. Contenido

Como el *valor_p* del 0.02356 es menor al nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Contenido y Fraude tienen una asociación estadísticamente significativa.

4.4.4. Continente

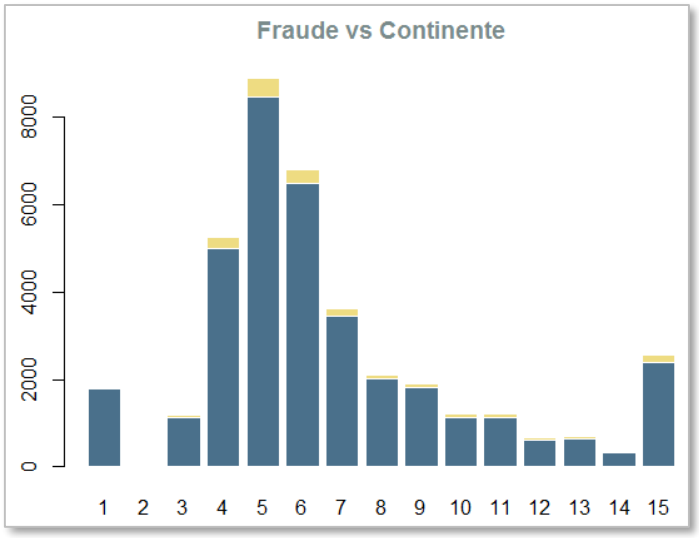


Imagen 7: Distribución variable Continente.
[Elaboración propia]

Continente		Fraude				
Categoría	Rango	No	Si	Total	Prop. No	Prop. Si
1	0	1.797	25	1.822	98,6%	1,4%
2	(0_20000]	42	1	43	97,7%	2,3%
3	(20000_40000]	1.131	45	1.176	96,2%	3,8%
4	(40000_60000]	4.987	256	5.243	95,1%	4,9%
5	(60000_80000]	8.455	425	8.880	95,2%	4,8%
6	(80000_100000]	6.488	300	6.788	95,6%	4,4%
7	(100000_120000]	3.451	158	3.609	95,6%	4,4%
8	(120000_140000]	2.006	83	2.089	96,0%	4,0%
9	(140000_160000]	1.807	91	1.898	95,2%	4,8%
10	(160000_180000]	1.131	77	1.208	93,6%	6,4%
11	(180000_200000]	1.138	72	1.210	94,0%	6,0%
12	(200000_220000]	623	37	660	94,4%	5,6%
13	(220000_240000]	650	48	698	93,1%	6,9%
14	(240000_250000]	337	18	355	94,9%	5,1%
15	>250000	2.399	162	2.561	93,7%	6,3%
Total		36.442	1.798	38.240	95%	5%

Tabla 4:Distribución variable Continente

En cuanto a la variable Continente se puede observar que la mayor proporción, en cuanto a rangos de continente asegurado, se encuentra entre 40.000 y 100.000, categorías 4,5 y 6 con proporciones sobre el total de 13.71%, 23.22% y 17.75% respectivamente.

También se observa que es una variable que está bien poblada ya que ni siquiera existe una categoría que se defina como "sin información", lo que ayuda a que las estimaciones en las que se tenga en cuenta esta variable generen resultados que se pueden interpretar y explicar.

En cuanto a la presencia de fraude dentro de las categorías se puede ver que es muy homogénea y se encuentra alrededor del 4.7%.

Al realizar el test Chi-cuadrado para validar si existe o no asociación entre las variables, se obtiene lo siguiente:

Pearson's Chi-squared test		
data: continente and Fraude		
X-squared = 88.223	df = 10	p-value = 8,212E-13

Resultado 3:Test Chi-Cuadrado: Fraude vs. Continente

El *valor_p* del 8.212E-13 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Continente y Fraude tienen una asociación estadísticamente significativa.

4.4.5. Tipo de mediador

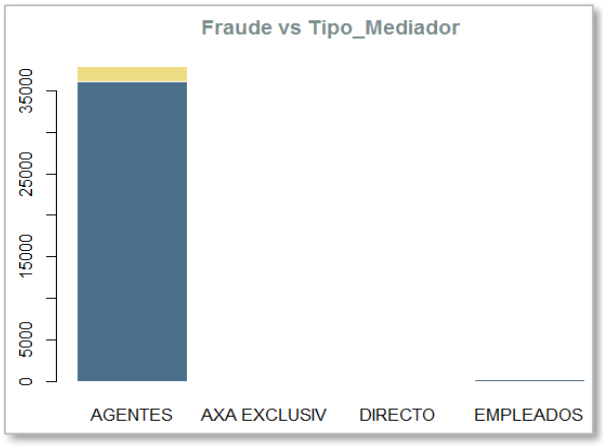


Imagen 8: Distribución variable Tipo Mediador.
[Elaboración propia]

	Fraude				
Tipo Mediador	No	Si	Total	Prop. No	Prop. Si
AGENTES	36.065	1.776	37.841	95,3%	4,7%
EXCLUSIVE	40	5	45	88,9%	11,1%
DIRECTO	95	14	109	87,2%	12,8%
EMPLEADOS	242	3	245	98,8%	1,2%
Total	36.442	1.798	38.240	95%	5%

Tabla 5:Distribución variable Tipo Mediador

Esta variable tiene claramente una categoría mayoritaria: **Agentes,** la cual contiene el 98.96% del total de datos. Sin embargo, cuando se analiza la proporción de fraude dentro de cada una de las categorías se puede observar que la categoría **Directo** presenta la mayor proporción de pólizas fraudulentas.

Al realizar el test Chi-cuadrado se obtienen los siguientes valores:

Pearson's Chi-squared test		
data: tipo de mediador and Fraude		
X-squared = 26,87	df = 3	p-value = 6,268E-6

Resultado 4:Test Chi-Cuadrado: Fraude vs. Tipo de mediador

El *valor_p* del 6.268E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Tipo de mediador y Fraude tienen una asociación estadísticamente significativa.

4.4.6. Forma de pago

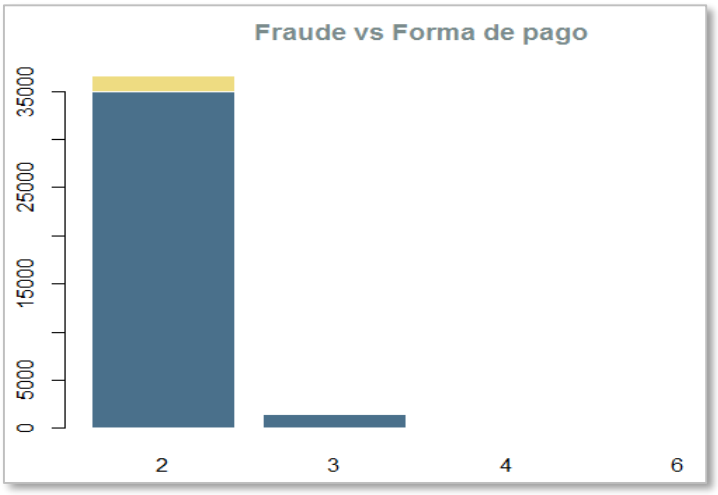


Imagen 9: Distribución variable Forma de pago.
[Elaboración propia]

Forma de Pago		Fraude				
Categoría	Descripción	No	Si	Total	Prop. No	Prop. Si
2	Anual	34.902	1.670	36.572	95,4%	4,6%
3	Semestral	1.478	116	1.594	92,7%	7,3%
4	Trimestral	59	12	71	83,1%	16,9%
6	Irregular	3	-	3	100,0%	0,0%
Total		36.442	1.798	38.240	95%	5%

Tabla 6:Distribución variable Forma de pago

Se puede observar que, aunque la forma de pago Anual es la más utilizada para el pago de las primas, la mayor proporción de fraude lo tienen aquellas en las que se realiza un pago de forma trimestral (16.9% de los datos de esta categoría son clasificados como Fraude, aunque del total de datos de la variable, esta categoría solo representa el 0.19%).

Al realizar el test Chi-cuadrado se obtiene lo siguiente:

Pearson's Chi-squared test		
data: forma de pago and Fraude		
X-squared = 48,825	df = 3	p-value = 1,421E-10

Resultado 5:Test Chi-Cuadrado: Fraude vs. Forma de pago

El $valor\_p$ del 1.421E-10 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Forma de pago y Fraude tienen una asociación estadísticamente significativa.

4.4.7. Nota de bureau

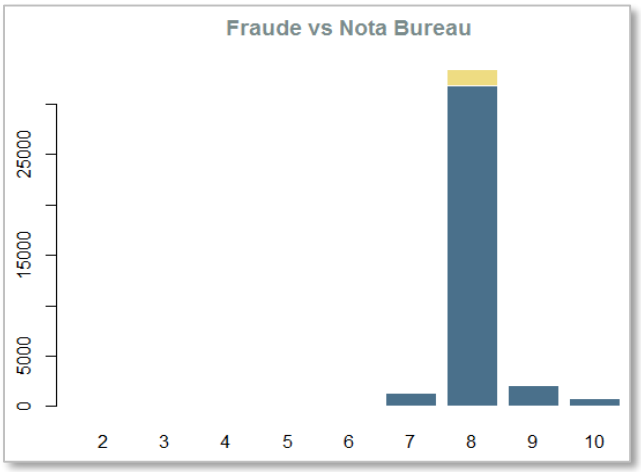


Imagen 10: Distribución variable Nota bureau.
[Elaboración propia]

Nota Bureau		Fraude				
Categoría	Forma de Pago	No	Si	Total	Prop. No	Prop. Si
2	N.I. Leve	7	-	7	100,0%	0,0%
3	N.I. Moderado	79	8	87	90,8%	9,2%
4	N.I. Importante	142	10	152	93,4%	6,6%
5	N.I. Crítico	45	9	54	83,3%	16,7%
6	N.I. Muy Crítico	4	-	4	100,0%	0,0%
7	Sin Info (Experian)	1.388	82	1.470	94,4%	5,6%
8	No está en la lista	31.875	1.557	33.432	95,3%	4,7%
9	Genérico	2.091	75	2.166	96,5%	3,5%
10	Sin datos (ESB)	811	57	868	93,4%	6,6%
Total		36.442	1.798	38.240	95%	5%

Tabla 7:Distribución variable Nota Bureau

En esta variable la categoría mayoritaria es la "**8. No está en la lista",** sin embargo, en términos de proporción de fraude y no fraude la categoría que cuenta con la mayor proporción en Fraude es la **"5. N.I. Crítico",** con un 16.7%.

Al realizar el test Chi-cuadrado se obtienen los siguientes valores:

Pearson's Chi-squared test		
data: nota bureau and Fraude		
X-squared = 39,741	df = 8	p-value = 3,58E-6

Resultado 6:Test Chi-Cuadrado: Fraude vs. Nota Bureau

El $valor\_p$ del 3.58E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Nota bureau y Fraude tienen una asociación estadísticamente significativa.

4.4.8. Nota global

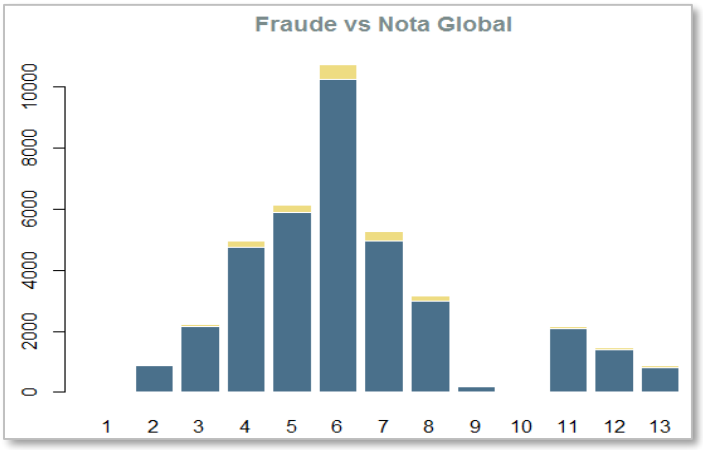


Imagen 11: Distribución variable Nota Global.
[Elaboración propia]

Nota Global		Fraude				
Categoría	Forma de Pago	No	Si	Total	Prop. No	Prop. Si
1	Sin Clasificación	67	2	69	97,1%	2,9%
2	Riesgo muy bajo	873	34	907	96,3%	3,7%
3	Riesgo muy bajo=bajo	2.142	95	2.237	95,8%	4,2%
4	Riesgo bajo	4.746	200	4.946	96,0%	4,0%
5	Riesgo bajo=medio	5.881	258	6.139	95,8%	4,2%
6	Riesgo medio	10.227	489	10.716	95,4%	4,6%
7	Riesgo medio=alto	4.970	296	5.266	94,4%	5,6%
8	Riesgo alto	2.976	182	3.158	94,2%	5,8%
9	Riesgo alto=muy alto	206	24	230	89,6%	10,4%
10	Riesgo muy alto	64	4	68	94,1%	5,9%
11	Genérico	2.091	75	2.166	96,5%	3,5%
12	Sin info (Experian)	1.388	82	1.470	94,4%	5,6%
13	Sin datos (ESB)	811	57	868	93,4%	6,6%
Total		36.442	1.798	38.240	95%	5%

Tabla 8:Distribución variable Nota global

Es interesante ver como en esta variable la mayor proporción de los datos se encuentra concentrada en la categoría en la que los tomadores de las pólizas están clasificados como **Riesgo medio,** mientras que en la variable de Nota bureau la mayor proporción estaba concentrada en **No está en la lista.** Esto significa que, aunque para el sistema financiero el tomador de la póliza no representa ningún riesgo de impago, para la compañía, propiamente, si puede representarlo.

Se puede observar que la distribución del fraude en cada una de las categorías está alrededor del 4.7% si no se tiene en cuenta la categoría **"9. Riesgo alto = muy alto"** ya que es la que tiene la mayor proporción de fraude (10.4%).

Al realizar el test, se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: nota global and Fraude		
X-squared = 63.659	df = 12	p-value = 4,821E-9

Resultado 7:Test Chi-Cuadrado: Fraude vs. Nota Global

El *valor_p* del 4.821E-9 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Nota global y Fraude tienen una asociación estadísticamente significativa.

4.4.9. Perfil del mediador

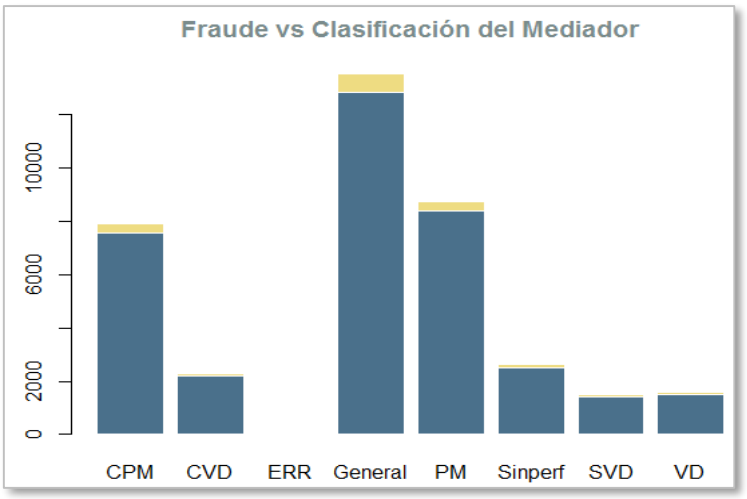


Imagen 12: Distribución variable Clasificación del mediador.
[Elaboración propia]

	Fraude				
Clasificación del Mediador	No	Si	Total	Prop. No	Prop. Si
PM	8.385	343	8.728	96,1%	3,9%
CPM	7.589	337	7.926	95,7%	4,3%
General	12.834	705	13.539	94,8%	5,2%
CVD	2.182	122	2.304	94,7%	5,3%
VD	1.521	85	1.606	94,7%	5,3%
SVD	1.409	94	1.503	93,7%	6,3%
Sinperf	2.520	112	2.632	95,7%	4,3%
ERR	2	-	2	100,0%	0,0%
Total	36.442	1.798	38.240	95%	5%

Tabla 9:Distribución variable Clasificación del mediador

En cuanto a la Clasificación del mediador, se puede observar que la mayor proporción de pólizas nuevas las generan los mediadores que están clasificados como General (representan un 35.41% del total), sin embargo, cuando se analiza la proporción de fraude en cada categoría (se visualiza más en la tabla que en la gráfica), se puede observar que la mayor proporción se encuentra en la categoría **SVD** (Súper Value Destroyer) (que es la categoría más baja –peor- de la clasificación). La mejor categoría dentro de esta variable es la categoría **PM** (Profit Maker)**,** y aunque representa un número significativo de pólizas respecto del total (22.82%), en cuanto a la proporción de fraude no lo es (3.9%).

Al realizar el test Chi-cuadrado para validar si existe o no asociación entre las variables, se obtiene lo siguiente:

Pearson's Chi-squared test		
data: perfil and Fraude		
X-squared = 35,32	df = 7	p-value = 9,738E-6

Resultado 8:Test Chi-Cuadrado: Fraude vs. Perfil

El *valor_p* del 9.738E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables Perfil del Mediador y Fraude tienen una asociación estadísticamente significativa.

4.4.10. Provincia

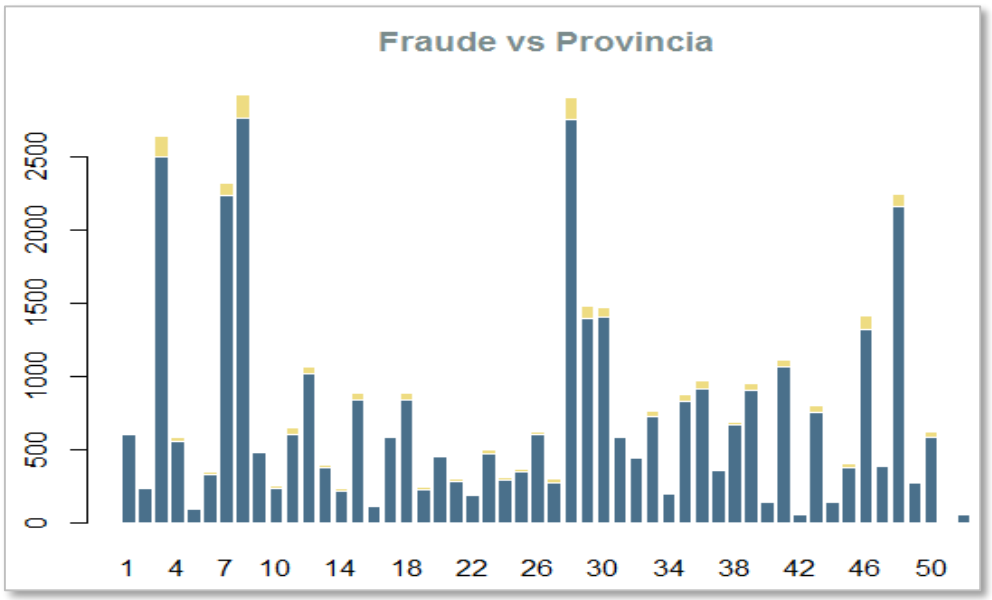


Imagen 13: Distribución variable Provincia.
[Elaboración propia]

En total se tienen 52 categorías dentro de esta variable, donde las más representativas en cuanto a número de pólizas nuevas son las provincias: 28-Madrid (7.58%), 8-Barcelona (7.64%), 3-Alicante (6.91%), 7-Baleares (6.07%) y la 48-Vizcaya (Bilbao) (5.88%). Sin embargo, al revisar las proporciones de fraude dentro de cada una de esas provincias se encuentra que las que sobresalen por tener una proporción de datos con fraude superior al 6% son (datos ordenados de mayor a menor en porcentaje de fraude):

Provincia	Nombre	No	Si	Total	Prop. No	Prop. Si
51	Ceuta	4	1	5	80,0%	20,0%
14	Córdoba	219	21	240	91,3%	8,8%
27	Lugo	279	25	304	91,8%	8,2%
16	Cuenca	116	10	126	92,1%	7,9%
19	Guadalajara	229	18	247	92,7%	7,3%
46	Valencia	1.320	98	1.418	93,1%	6,9%
10	Cáceres	237	17	254	93,3%	6,7%
11	Cádiz	608	42	650	93,5%	6,5%
36	Pontevedra	915	62	977	93,7%	6,3%
43	Tarragona	756	50	806	93,8%	6,2%
23	Jaén	476	31	507	93,9%	6,1%

Tabla 10:Distribución variable Provincia

De estas provincias la que más llama la atención es la **51-Ceuta**, ya que de 5 pólizas nuevas que tiene para los dos años estudiados, **4 de ellas han resultado identificadas como fraude** al momento de reclamación de algún siniestro.

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: provincia and Fraude		
X-squared = 162,05	df = 51	p-value = 1,721E-13

Resultado 9:Test Chi-Cuadrado: Fraude vs. Provincia

El $valor\_p$ del 1.721E-13 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que las variables provincia y Fraude tienen una asociación estadísticamente significativa.

4.4.11. Superficie

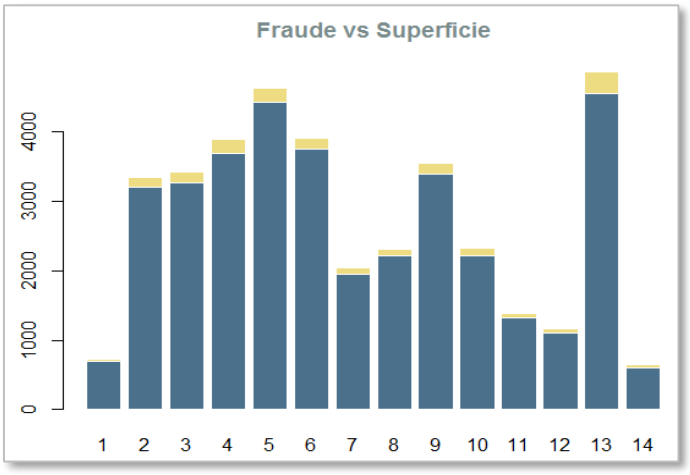


Imagen 14: Distribución variable Superficie.
[Elaboración propia]

Superficie		Fraude				
Categoría	Rango (m^2)	No	Si	Total	Prop. No	Prop. Si
1	(0_40]	699	22	721	96,9%	3,1%
2	(40_60]	3.203	143	3.346	95,7%	4,3%
3	(60_70]	3.269	150	3.419	95,6%	4,4%
4	(70_80]	3.694	195	3.889	95,0%	5,0%
5	(80_90]	4.428	210	4.638	95,5%	4,5%
6	(90_100]	3.756	159	3.915	95,9%	4,1%
7	(100_110]	1.948	92	2.040	95,5%	4,5%
8	(110_120]	2.223	94	2.317	95,9%	4,1%
9	(120_140]	3.390	155	3.545	95,6%	4,4%
10	(140_160]	2.223	112	2.335	95,2%	4,8%
11	(160_180]	1.332	56	1.388	96,0%	4,0%
12	(180_200]	1.110	64	1.174	94,5%	5,5%
13	>200	4.556	308	4.864	93,7%	6,3%
14	9999	611	38	649	94,1%	5,9%
Total		36.442	1.798	38.240	95%	5%

Tabla 11:Distribución variable Superficie

La distribución de los datos dentro de las categorías de esta variable está muy homogénea (alrededor de 7,14%), aunque sobresale la categoría con pólizas sobre viviendas con **superficie >200 m²,** ya que no solo es la que mayor proporción de datos tiene sobre el total (12.72%), sino que es en la que mayor porcentaje de fraude se ha detectado (6.3%).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Superficie and Fraude		
X-squared = 48,109	df = 13	p-value = 6,27E-6

Resultado 10:Test Chi-Cuadrado: Fraude vs. Superficie

El *valor_p* del 1. 6.27E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que los metros cuadrados que tenga la vivienda y la variable Fraude tienen una asociación estadísticamente significativa.

4.4.12.Territorial

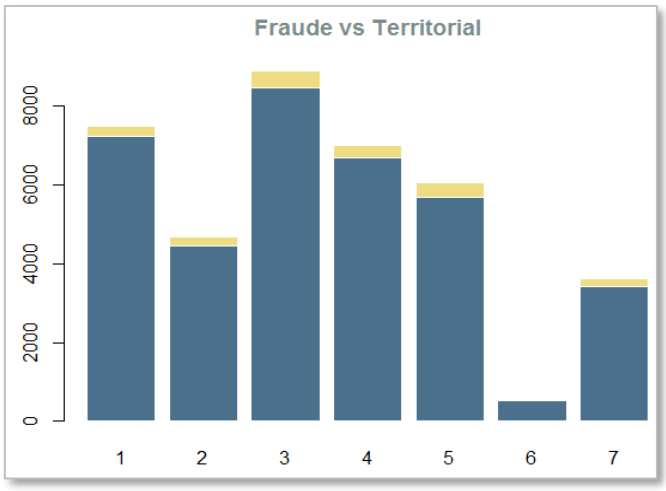


Imagen 15: Distribución variable Territorial.
[Elaboración propia]

		Fraude				
Categoría	Territorial	No	Si	Total	Prop. No	Prop. Si
1	Norte	7.217	262	7.479	96,5%	3,5%
2	Este	4.458	232	4.690	95,1%	4,9%
3	Levante	8.456	418	8.874	95,3%	4,7%
4	Centro	6.683	310	6.993	95,6%	4,4%
5	Sur	5.685	356	6.041	94,1%	5,9%
6	Servicios Centrales	516	34	550	93,8%	6,2%
7	Oeste	3.427	186	3.613	94,9%	5,1%
Total		36.442	1.798	38.240	95%	5%

Tabla 12:Distribución variable Territorial

Las territoriales en las que más contratan seguros de hogar son Levante (23.21%) y Norte (19.56%). Sin embargo, en cuanto a la distribución del fraude, se puede ver que la territorial de Servicios Centrales es la que mayor concentración tiene (6.2%), seguido de la territorial Sur (5.9%).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Territorial and Fraude		
X-squared = 49,165	df = 6	p-value = 6,909E-9

Resultado 11:Test Chi-Cuadrado: Fraude vs. Territorial

El *valor_p* del 1.721E-13 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que la zona geográfica donde opera el agente que realizo la contratación de la póliza y a variable Fraude tienen una asociación estadísticamente significativa.

4.4.13. Tipo de vivienda

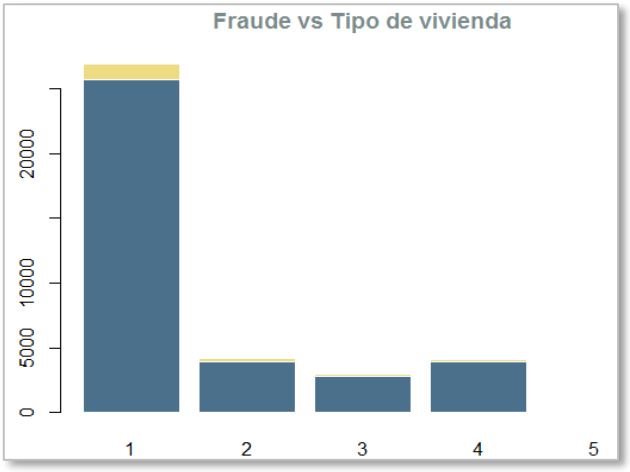


Imagen 16: Distribución variable Tipo de vivienda.
[Elaboración propia]

Tipo de Vivienda		Fraude				
Categoría	Rango (m^2)	No	Si	Total	Prop. No	Prop. Si
1	Piso	25.679	1.187	26.866	95,6%	4,4%
2	Chalet Adosado	3.980	220	4.200	94,8%	5,2%
3	Chalet Independiente	2.839	163	3.002	94,6%	5,4%
4	Casa Tradicional	3.912	227	4.139	94,5%	5,5%
5	Otras viviendas	32	1	33	97,0%	3,0%
Total		36.442	1.798	38.240	95%	5%

Tabla 13:Distribución variable Vivienda

De la gráfica se puede observar que la mayor cantidad de pólizas que se contratan en el seguro de vivienda son para asegurar pisos/apartamentos (70.26%), sin embargo, en cuanto a proporción de fraude, es más representativa la casa

tradicional (aunque la diferencia solo es de 0.3% respecto al promedio de las categorías en esta variable)

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Tipo de vivienda and Fraude		
X-squared = 16,931	df = 4	p-value =0.001994

Resultado 12:Test Chi-Cuadrado: Fraude vs. Tipo de vivienda

El *valor_p* del 0.001994 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que el tipo de vivienda y la variable Fraude tienen una asociación estadísticamente significativa.

4.4.14. Ubicación

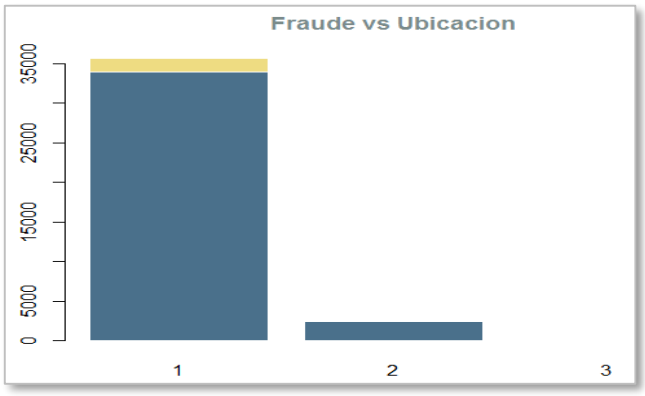


Imagen 17: Distribución variable Ubicación.
[Elaboración propia]

		Fraude				
Categoría	Ubicación	No	Si	Total	Prop. No	Prop. Si
1	Casco urbano	33.943	1.702	35.645	95,2%	4,8%
2	Urbanización	2.464	96	2.560	96,3%	3,8%
3	Despoblado	35	-	35	100,0%	0,0%
Total		36.442	1.798	38.240	95%	5%

Tabla 14:Distribución variable Ubicación

Tanto de la gráfica como de la tabla, se puede observar que la mayor cantidad de viviendas aseguradas se encuentran en el casco urbano (93.21%), además también es la ubicación en la que mayor proporción de fraude se da (4.8%).
Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Ubicación and Fraude		
X-squared = 7,3272	df = 2	p-value = 0.02564

Resultado 13:Test Chi-Cuadrado: Fraude vs. Ubicación

El $valor\_p$ del 0.02564 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que la ubicación de la vivienda y la variable Fraude tienen una asociación estadísticamente significativa.

4.4.15. Tipo de usuario

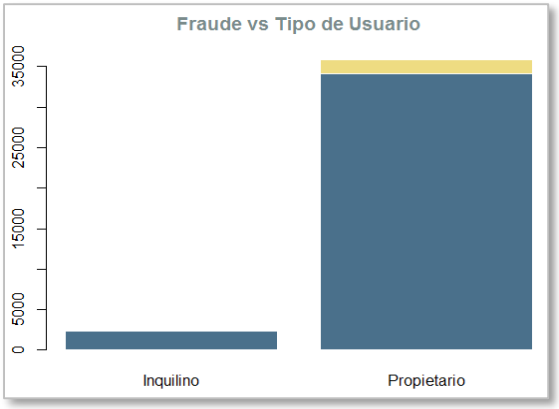


Imagen 18: Distribución variable Tipo de usuario.
[Elaboración propia]

	Fraude				
Tipo de Usuario	No	Si	Total	Prop. No	Prop. Si
Inquilino	2.333	68	2.401	97,2%	2,8%
Propietario	34.109	1.730	35.839	95,2%	4,8%
Total	36.442	1.798	38.240	95%	5%

Tabla 15:Distribución variable Tipo de usuario

Se observa que el tipo de usuario que más contrata seguros de hogar son los propios dueños de las viviendas (93.72% del total) y además que es en esta misma categoría en donde se presenta la mayor proporción de fraude dentro de la variable (4.8%)

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Tipo de usuario and Fraude		
X-squared = 19,545	df = 1	p-value = 9,828E-6

Resultado 14:Test Chi-Cuadrado: Fraude vs. Tipo de usuario

El *valor_p* del 9.828E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que el tipo de usuario y la variable Fraude tienen una asociación estadísticamente significativa.

4.4.16. Tipo de Uso

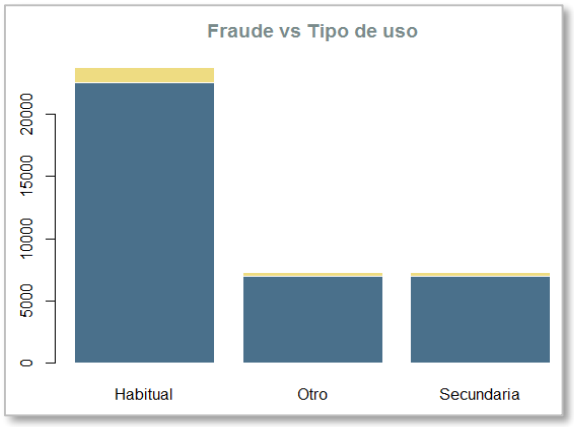


Imagen 19: Distribución variable Tipo de uso.
[Elaboración propia]

	Fraude				
Tipo de Uso	No	Si	Total	Prop. No	Prop. Si
Habitual	22.496	1.206	23.702	94,9%	5,1%
Otro	6.951	328	7.279	95,5%	4,5%
Secundaria	6.995	264	7.259	96,4%	3,6%
Total	36.442	1.798	38.240	95%	5%

Tabla 16:Distribución variable Tipo de uso

Se observa que la categoría de Tipo de Uso-Habitual no solo presenta el mayor número de datos (61.98%), sino también la que presenta la mayor proporción de fraude dentro de la categoría (5.1%).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Tipo de uso and Fraude		
X-squared = 26,891	df = 2	p-value = 1,447E-6

Resultado 15:Test Chi-Cuadrado: Fraude vs. Tipo de uso

El *valor_p* del 1.447E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que el tipo de uso y la variable Fraude tienen una asociación estadísticamente significativa.

4.4.17. Rehabilitada

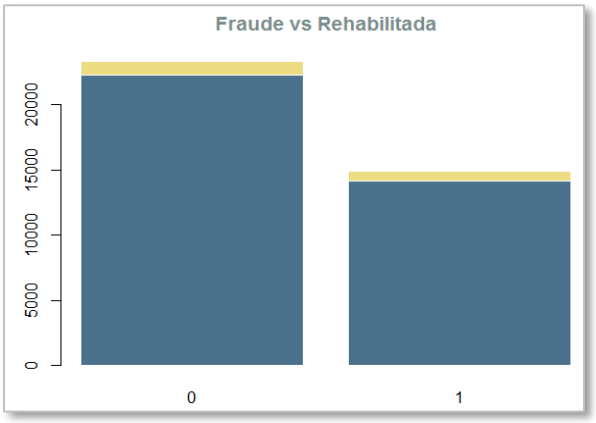


Imagen 20:Distribución variable Rehabilitada
[Elaboración propia]

	Fraude				
Rehabilitada	No	Si	Total	Prop. No	Prop. Si
No (0)	22.259	1.043	23.302	95,5%	4,5%
SI (1)	14.183	755	14.938	94,9%	5,1%
Total	36.442	1.798	38.240	95%	5%

Tabla 17:Distribución variable Rehabilitada

Tanto de la gráfica como de la tabla, se puede observar que la mayor cantidad de viviendas aseguradas no han sido rehabilitadas (60.9%), y que la mayor proporción de fraude se encuentra en las que sí lo han sido (5.1 %).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Rehabilitada and Fraude		
X-squared = 6,6634	df = 1	p-value = 0,009841

Resultado 16:Test Chi-Cuadrado: Fraude vs. Rehabilitada

El $valor\_p$ del 0.009841 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que si la vivienda ha sido rehabilitada o no y la variable Fraude, tienen una asociación estadísticamente significativa.

4.4.18. Edad del tomador

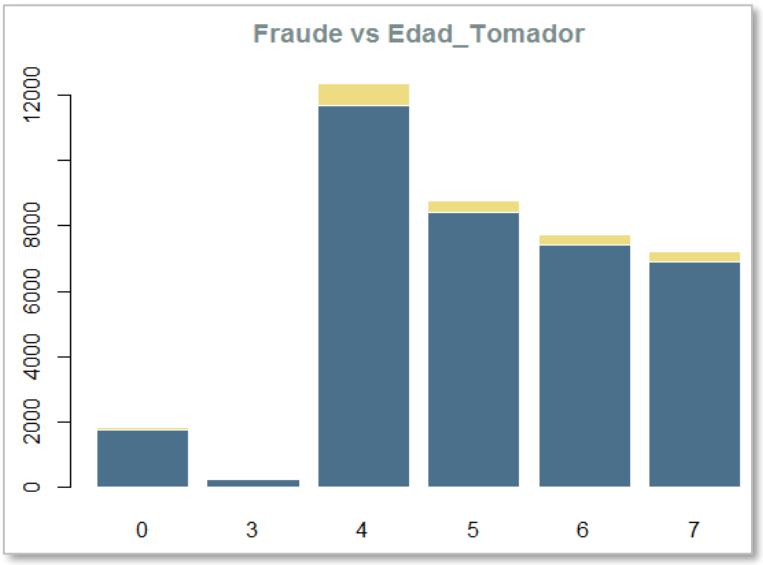


Imagen 21:Distribución variable Edad del tomador
[Elaboración propia]

		Fraude				
Categoría	Edad del tomador	No	Si	Total	Prop. No	Prop. Si
0	Sin informar	1.788	59	1.847	96,8%	3,2%
3	[14,26) años	262	14	276	94,9%	5,1%
4	[26,50) años	11.669	680	12.349	94,5%	5,5%
5	[50,60) años	8.406	389	8.795	95,6%	4,4%
6	[60,70) años	7.408	342	7.750	95,6%	4,4%
7	>=70 años	6.909	314	7.223	95,7%	4,3%
Total		36.442	1.798	38.240	95%	5%

Tabla 18:Distribución variable Edad del tomador

Esta variable muestra que la mayor concentración de personas que contratan un seguro de hogar está entre las edades de 26 a 50 años (32.29% del total, aunque hay que considerar también que el rango es bastante amplio) y que además es en este rango en el que se encuentra la mayor proporción de fraude dentro de las categorías de la variable (5.5%).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Edad del tomador and Fraude		
X-squared = 32.294	df = 5	p-value = 5,194E-6

Resultado 17:Test Chi-Cuadrado: Fraude vs. Edad del tomador

El *valor_p* del 5.194E-6 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que la edad del tomador de la póliza y la variable Fraude, tienen una asociación estadísticamente significativa

4.4.19. Hipoteca

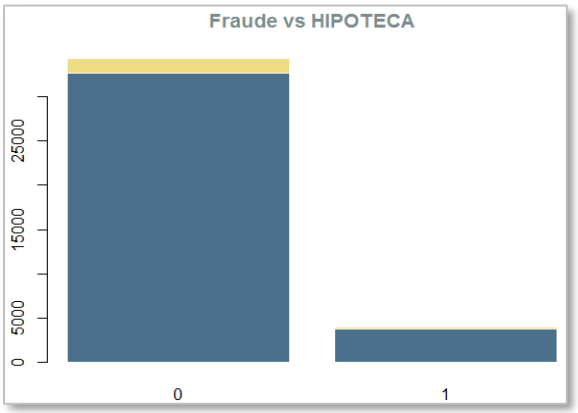


Imagen 22:Distribución variable Hipoteca
[Elaboración propia]

	Fraude				
Hipoteca	No	Si	Total	Prop. No	Prop. Si
No (0)	32.676	1.579	34.255	95,4%	4,6%
SI (1)	3.766	219	3.985	94,5%	5,5%
Total	36.442	1.798	38.240	95%	5%

Tabla 19:Distribución variable Hipoteca

Esta variable muestra que, en su mayoría, las viviendas aseguradas no están hipotecadas (89.58%), sin embargo, la mayor proporción de fraude se encuentra en las pólizas correspondientes a viviendas que sí están hipotecadas (5.5%).

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Hipoteca and Fraude		
X-squared = 6.0585	df = 1	p-value = 0.01384

Resultado 18:Test Chi-Cuadrado: Fraude vs. Hipoteca

El *valor_p* del 0.01384 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que si la vivienda tiene o no hipoteca y la variable Fraude, tienen una asociación estadísticamente significativa

4.4.20. Valor del cliente

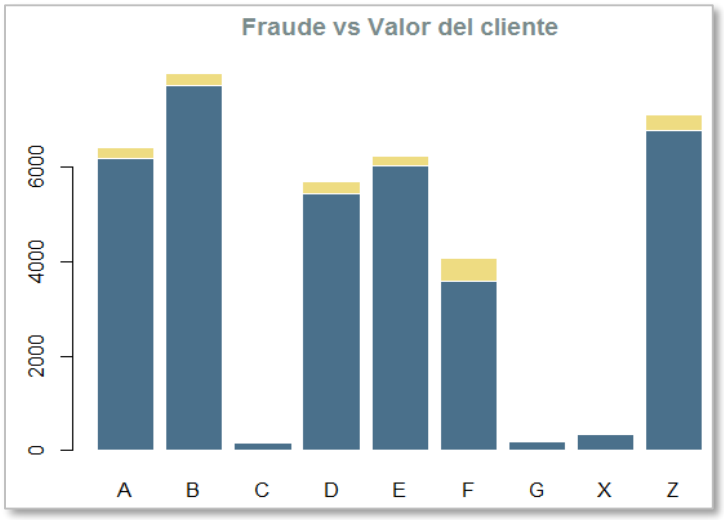


Imagen 23: Distribución variable Valor del cliente
[Elaboración propia]

	Fraude				
Valor del cliente	No	Si	Total	Prop. No	Prop. Si
A	6.198	219	6.417	96,6%	3,4%
B	7.718	262	7.980	96,7%	3,3%
C	164	7	171	95,9%	4,1%
D	5.437	264	5.701	95,4%	4,6%
E	6.029	200	6.229	96,8%	3,2%
F	3.588	477	4.065	88,3%	11,7%
G	188	2	190	98,9%	1,1%
X	353	15	368	95,9%	4,1%
Z	6.767	352	7.119	95,1%	4,9%
Total	36.442	1.798	38.240	95%	5%

Tabla 20:Distribución variable Valor del cliente

Tanto de la gráfica como de la tabla se puede observar que el porcentaje de fraude en las categorías es muy homogéneo excepto en la categoría de "valor del cliente: F" en donde se observa una gran diferencia de pólizas con fraude vs. las que tienen las demás categorías (11.7% vs. el promedio de la variable de 4.5%).

Es importante resaltar que las categorías de esta variable están ordenadas de acuerdo a la importancia que tiene el cliente para toda la compañía, y el orden es descendente, es decir, los clientes más importantes se encuentran en la categoría A y los menos importantes en la Z.

Al realizar la prueba Chi-cuadrado para validar la hipótesis de independencia se obtienen los siguientes resultados:

Pearson's Chi-squared test		
data: Valor del cliente and Fraude		
X-squared = 546.32	df = 8	p-value < 2,2E-16

Resultado 19:Test Chi-Cuadrado: Fraude vs. Valor del cliente

El valor_p <2.2E-16 indica que, bajo un nivel de significatividad del 0.05, se rechaza la hipótesis nula, por lo que se concluye que el valor que le ha asignado la compañía al tomador de la póliza y la variable Fraude, tienen una asociación estadísticamente significativa

4.5. Resumen test Chi-cuadrado

Para tener una visión más general de los resultados obtenidos con este test se construye la siguiente tabla, la cual está construida en forma descendente de la columna p-value.

	Pearson's Chi-squared test		
Variable	X-squared	df	p-value
Antigüedad	15,909	10	0,1023
Ubicación	7,327	2	0,02564
Contenido	20,664	10	0,02356
Hipoteca	6,059	1	0,01384
Rehabilitada	6,663	1	0,009841
Tipo de vivienda	16,931	4	0,001994
Tipo de usuario	19,545	1	9,83E-06
Perfil del mediador	35,320	7	9,74E-06
Superficie	48,109	13	6,27E-06
Tipo de Mediador	26,870	3	6,27E-06
Edad del tomador	32,294	5	5,19E-06
Nota Bureau	39,741	8	3,58E-06
Tipo de uso	26,891	2	1,45E-06
Territorial	49,165	6	6,909E-09
Nota Global	63,659	12	4,821E-09
Forma de pago	48,825	3	1,421E-10
Continente	88,223	10	8,212E-13
Provincia	162,050	51	1,721E-13
Valor del cliente	546,320	8	2,2E-16

Tabla 21: Resumen resultados Test Chi-Cuadrado

Con esta tabla se puede ver que la única variable con la que se obtiene un *valor_p* superior al nivel de significatividad del 0,05 es **Antigüedad de la vivienda,** por lo tanto (y como se mencionó antes), sobre la relación de estas dos variables no se puede concluir que exista algún tipo de asociación.

Como uno de los objetivos principales de este trabajo es implementar un modelo GLM que ayude en la predicción (detección) de las posibles futuras pólizas fraudulentas, con el resultado anterior se podría considerar excluir desde este

momento la variable antigüedad del modelo GLM como variable predictora. Sin embargo, se incluirá para que los modelos de selección de variables (en este caso Boruta y método de extracción AIC) realicen la validación con el total de las variables y en sus procesos la excluyan o, por qué no, la incluyan si encuentran relaciones entre ésta y otras variables que puedan hacerla significativa en el GLM.

5. DIVISIÓN DEL CONJUNTO DE DATOS

Dentro de los algoritmos de *machine learning*, los modelos GLM están categorizados dentro de los algoritmos de clasificación que operan bajo el aprendizaje automático supervisado, esto se debe a que cumple con la característica de que los datos de entrada son conocidos y etiquetados de tal forma que se pueda deducir un determinado patrón o función a partir de estos. (Daymon, 2018)

En general, cuando se utilizan este tipo de modelos, lo que se busca es que el algoritmo utilizado se ajuste bien a los datos pasados y a su vez realice una predicción con buena exactitud, sin embargo, por la naturaleza asimétrica de los datos que se tienen para este análisis, las exigencias que se realizan al modelo son diferentes ya que no dependen solamente de la "exactitud" (ya que esta medida asigna la misma importancia a predecir bien en ambas categorías de la variable respuesta), sino de todas las medidas en conjunto.

Lo que sí es común en todos los modelos, independiente de si tienen datos asimétricos o no, es realizar una partición de los datos en dos conjuntos: uno de entrenamiento y otro de validación:

- Conjunto de datos de entrenamiento: como su nombre lo indica, se utiliza para **entrenar** los algoritmos y ajustar los hiperparámetros.
- Conjunto de datos de validación: después de obtener un modelo final, se utilizan estos datos para estimar el error de predicción.

Las formas más comunes de hacer la partición del conjunto de datos para entrenamiento y validación son, respectivamente:

a) 60% - 40%
b) 70% - 30%
c) 80% - 20%

La división utilizada en este trabajo es la **b) 70% - 30%.** Esto se decide teniendo en cuenta que en la opción a) se dejan muchos datos para pruebas pudiendo evitar un buen ajuste de los parámetros del modelo y que en la opción c) se tienen demasiados datos para el entrenamiento pudiendo afectar la adecuada evaluación del rendimiento predictivo.

Ahora, para controlar la asimetría, se utiliza **un muestreo estratificado** para la partición, esto garantiza una representación equilibrada de la distribución de la variable respuesta en ambos conjuntos. La forma de hacer esto en R es con el uso

del paquete *rsample* (Kuhn , Chow , & Wickham, 2019), donde se debe especificar la variable respuesta para estratificar (que en este caso es la variable Fraude). Al ejecutarlo se obtienen las siguientes proporciones:

	No Fraude	Fraude
Conjunto de entrenamiento	95,23%	4,77%
Conjunto de validación	95,46%	4,54%
Total Datos	**95,30%**	**4,70%**

Resultado 20: Distribución de la variable fraude en conjuntos de entrenamiento y validación

Después de tener ambos conjuntos definidos, se procede al análisis de significatividad de las variables para determinar cuáles de ellas son relevantes a la hora de realizar la predicción con el GLM.

6. SELECCIÓN DE LAS VARIABLES A UTILIZAR EN EL GLM A PARTIR DE LOS DATOS DEL CONJUNTO DE ENTRENAMIENTO

Para la selección de variables que se deberían incluir en el GLM se comparan dos algoritmos de selección pertenecientes a los denominados *procedimientos stepwise* (Guerrero, 2016), uno de ellos es el algoritmo Boruta y el otro el procedimiento de extracción AIC.

6.1. Algoritmo Boruta

Para dar una idea de cómo funciona este algoritmo de clasificación, se muestra la definición hecha por (Guerrero, 2016):

> *Boruta es un método que utiliza random forest como algoritmo subyacente. La idea es generar en cada iteración una serie de variables sombra a partir de los predictores, copiando cada uno de ellos y permutando entre sí los elementos de cada nueva columna. Se ajusta un modelo por random forest y se calculan las importancias relativas de cada variable. Si una variable sistemáticamente queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo será dudosa y por tanto se elimina. El proceso continúa hasta que todas las variables son aceptadas, rechazadas o se alcanza un número de iteraciones límite.*

En esencia, *Boruta* busca capturar todas las características importantes e interesantes que pueda tener un conjunto de datos respecto a una variable respuesta. Cuando el algoritmo arroja los resultados, clasifica las variables predictoras en 3 niveles (Pathak, 2018):

- Sin Importancia
- Tentativa de ser importante
- Importante

La categoría de "Tentativa" significa que las características de las variables clasificadas así, tienen una importancia que está muy cerca de las características de las variables importantes, pero no suficientes para quedar en dicha categoría, por lo que el algoritmo lo deja clasificado de tal forma que sea el analista quien decida si la incorpora o no dentro de su modelo.

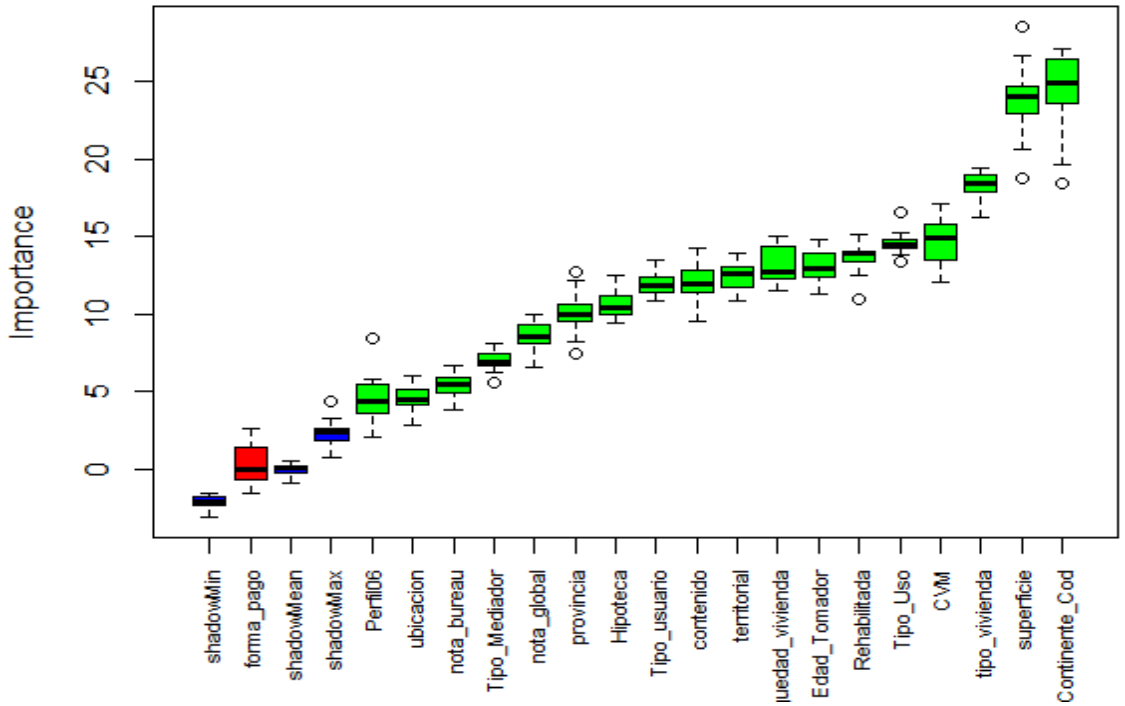
Aplicando el algoritmo a la base de datos, se obtiene que la clasificación de las 19 variables predictoras en los niveles de respuesta es la siguiente:

- **Sin Importancia:** sólo una de las variables es considerada como irrelevante. Esta variable es "forma de pago"

- **Tentativa de ser importante**: ninguna de las variables fue clasificada dentro de este nivel.

- **Importante:** 18 de las 19 variables se consideran relevantes para el análisis (todas excepto "forma de pago").

Una ventaja que tiene este algoritmo es que permite ver el resultado de forma gráfica a través de un gráfico de boxplots, en donde cada variable es representada por una caja y los colores representan cada uno de los niveles: sin importancia en rojo, tentativa de ser importante en amarillo e importante en verde[8], además, hay que tener en cuenta que el eje representa la importancia de cada variable dentro del conjunto de datos analizado.



Resultado 21: Gráfico Boruta de significatividad de las variables

En este caso se puede observar que "forma de pago" sale de color rojo (al ser clasificada dentro del nivel: sin importancia) y que el resto de variables son consideradas importantes para el análisis (están de color verde), siendo "**Continente", "Superficie" y "Tipo de vivienda**", las variables más importantes del conjunto de datos.

[8] El gráfico también muestra unos boxplot de color azul los cuales corresponden a la puntuación Z mínimo, promedio y máximo de una característica de sombra, creada por el propio algoritmo.

La siguiente es la tabla de los parámetros del gráfico:

```
> print(borutadf)
                      meanImp   medianImp    minImp    maxImp  normHits  decision
antiguedad_vivienda 13.1025001 12.74041686 11.453390 14.985609 1.0000000 Confirmed
contenido           12.0521388 11.98433663  9.479713 14.306522 1.0000000 Confirmed
Continente_Cod      24.4270328 24.98834711 18.456563 27.129764 1.0000000 Confirmed
Tipo_Mediador        7.0170239  6.92428218  5.513251  8.084107 1.0000000 Confirmed
forma_pago           0.2720364 -0.01967505 -1.596192  2.618774 0.1052632  Rejected
nota_bureau          5.4195776  5.49828599  3.769019  6.651379 1.0000000 Confirmed
nota_global          8.5896216  8.56839627  6.523260  9.920007 1.0000000 Confirmed
Perfil06             4.5348998  4.34638090  2.044645  8.384270 0.9473684 Confirmed
provincia           10.0880900  9.97487682  7.390470 12.689774 1.0000000 Confirmed
superficie          23.8477860 24.02558844 18.793110 28.600718 1.0000000 Confirmed
territorial         12.4074124 12.62583635 10.817538 13.928445 1.0000000 Confirmed
tipo_vivienda       18.3010772 18.47754318 16.275889 19.470732 1.0000000 Confirmed
ubicacion            4.5785000  4.49068081  2.827231  5.999293 1.0000000 Confirmed
Tipo_usuario        11.9897855 11.84026758 10.865088 13.538484 1.0000000 Confirmed
Tipo_Uso            14.5948795 14.44735174 13.331451 16.529843 1.0000000 Confirmed
Rehabilitada        13.7143655 13.90730557 10.966125 15.124641 1.0000000 Confirmed
Edad_Tomador        13.1352949 12.91037769 11.333810 14.860194 1.0000000 Confirmed
Hipoteca            10.6085321 10.43662165  9.448007 12.492732 1.0000000 Confirmed
CVM                 14.7019830 14.88395203 12.086139 17.131368 1.0000000 Confirmed
```

Resultado 22:Significatividad de las variables utilizando algoritmo Boruta

6.2. Método de extracción AIC

Otra forma de realizar la selección de las variables a tener en cuenta en la elaboración del modelo de predicción es utilizar el valor AIC como referente, para esto se utiliza la función *step* del paquete *stats* de R (R Core Team, 2019). Este algoritmo hace parte de los procedimientos *stepwise* ejecutados de forma descendiente (Guerrero, 2016), en particular, este algoritmo parte del conjunto de todas las variables predictoras y determina en cada paso (cada ejecución del GLM), la variable que menos aporta al modelo en función del AIC, y la elimina; este proceso de eliminación lo realiza hasta que eliminar alguna de las variables ya no disminuye el AIC del modelo [9].

La instrucción de R utilizada para indicar el modelo GLM aplicado a la variable fraude utilizando el resto de variables como predictoras es la siguiente:

```
M1_i1 <- glm (Fraude~., family = binomial (link = logit), data= train)
```

[9] En el anexo 14.2 se muestra la secuencia de eliminación de variables seguida por el algoritmo.

Aquí se debe utilizar como función de enlace la función logit con la familia binomial, ya que el objetivo es hallar la relación que existe entre un conjunto de variables categóricas con una variable respuesta dicotómica tal que 1: éxito (Fraude) y 0: Fracaso (No fraude).
Luego, para obtener el modelo óptimo en función del AIC, se ejecuta el siguiente código:

```
MFinal_i1 <- step (M1_i1, test = "Chisq")
```

Y se obtiene que las variables con las que se obtienen mejores resultados (menor AIC), es:

```
Step:  AIC=9811.67
Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador +
    forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso +
    Edad_Tomador + CVM

                     Df Deviance     AIC    LRT               Pr(>Chi)
<none>                    9689.7  9811.7
- antiguedad_vivienda 10   9711.4  9813.4  21.68                0.01681 *
- forma_pago           3   9697.4  9813.4   7.74                0.05159 .
- Perfil06             7   9707.7  9815.7  18.07                0.01166 *
- Tipo_Mediador        3   9700.1  9816.1  10.43                0.01524 *
- ubicacion            2   9698.4  9816.4   8.69                0.01296 *
- Tipo_Uso             2   9708.9  9826.9  19.21       0.00006731953523 ***
- territorial          6   9720.4  9830.4  30.71       0.00002883917641 ***
- Edad_Tomador         5   9725.3  9837.3  35.59       0.00000114750731 ***
- Continente_Cod      14   9769.5  9863.5  79.87       0.00000000002993 ***
- CVM                  8  10008.3 10114.3 318.66 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado 23: Significatividad de las variables seleccionadas por el modelo de reducción del AIC

Como resultado se obtiene que todas las variables que quedan en el modelo final son significativas (excepto “forma de pago” que supera por muy poco el nivel de significatividad), siendo las más relevantes: **Tipo de uso, Territorial, Edad del tomador, Continente y CVM.**

6.3. Comparación de los modelos y selección del modelo final.

Para comparar el modelo inicial (donde las 19 variables son predictoras) con los dos modelos utilizados para la selección de variables, se genera la tabla Anova para cada uno de los 3 modelos (todos sobre el conjunto de entrenamiento):

Modelo inicial (M1_i1).

M1_i1 <- glm (Fraude~., family = binomial (link = logit), data= train)

```
> Anova(M1_i1)
Analysis of Deviance Table (Type II tests)

Response: Fraude
                   LR Chisq Df             Pr(>Chisq)
antiguedad_vivienda   22.140 10              0.0144083 *
contenido              9.792 10              0.4589124
Continente_Cod        30.324 14              0.0068849 **
Tipo_Mediador         10.346  3              0.0158434 *
forma_pago             3.639  3              0.3031503
nota_bureau           10.002  5              0.0751867 .
nota_global           19.865  9              0.0187672 *
Perfil06              14.504  7              0.0429144 *
provincia             91.164 51              0.0004676 ***
superficie             9.838 13              0.7070908
territorial            8.870  6              0.1810446
tipo_vivienda          6.758  4              0.1492545
ubicacion             12.474  2              0.0019561 **
Tipo_usuario           0.311  1              0.5768538
Tipo_Uso              14.792  2              0.0006136 ***
Rehabilitada           0.215  1              0.6428098
Edad_Tomador          27.624  5             0.00004311 ***
Hipoteca               1.847  1              0.1740788
CVM                  314.768  8 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado 24:ANOVA del modelo GLM utilizando todas las variables en el conjunto de entrenamiento

Modelo Boruta:

boruta.train <- Boruta (Fraude~., data = train, doTrace = 2)

```
> Anova(M1_Boruta)
Analysis of Deviance Table (Type II tests)

Response: Fraude
                   LR Chisq Df Pr(>Chisq)
antiguedad_vivienda   22.431 10  0.0130529 *
contenido              9.939 10  0.4458656
Continente_Cod        30.944 14  0.0056445 **
Tipo_Mediador         10.426  3  0.0152703 *
nota_bureau           10.270  5  0.0679304 .
nota_global           20.238  9  0.0164995 *
Perfil06              14.825  7  0.0383070 *
provincia             92.381 51  0.0003476 ***
superficie             9.917 13  0.7007339
territorial            8.888  6  0.1799515
tipo_vivienda          7.606  4  0.1071201
ubicacion             12.283  2  0.0021522 **
Tipo_usuario           0.272  1  0.6019263
Tipo_Uso              15.245  2  0.0004894 ***
Rehabilitada           0.207  1  0.6489014
Edad_Tomador          27.851  5  3.892e-05 ***
Hipoteca               2.061  1  0.1511555
CVM                  315.157  8  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado 25:ANOVA del modelo GLM Boruta

Modelo de extracción AIC:

MOD_OPT <- Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + CVM

Mod_FullData <- glm (MOD_OPT,family = binomial (link = logit), data= train)

```
> Anova(MFinal_i1)
Analysis of Deviance Table (Type II tests)

Response: Fraude
                  LR Chisq Df            Pr(>Chisq)
antiguedad_vivienda    21.68 10               0.01681 *
Continente_Cod         79.87 14      0.00000000002993 ***
Tipo_Mediador          10.43  3               0.01524 *
forma_pago              7.74  3               0.05159 .
Perfil06               18.07  7               0.01166 *
territorial            30.71  6      0.00002883917641 ***
ubicacion               8.69  2               0.01296 *
Tipo_Uso               19.21  2      0.00006731953523 ***
Edad_Tomador           35.59  5      0.00000114750731 ***
CVM                   318.66  8 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado 26: ANOVA del modelo óptimo hallado con reducción del AIC

Entre el modelo inicial y el modelo Boruta no hay mucha diferencia, se puede observar que incluyen variables que no presentan ninguna significatividad tales como superficie, territorial y tipo de vivienda. Ahora, comparando los resultados del modelo inicial y el modelo de extracción AIC, se puede observar que el proceso de selección por AIC, deja en el modelo final (como modelo óptimo) la mayor parte de las variables que en el modelo inicial son significativas: Continente, Tipo de uso, Edad y CVM. Es interesante ver que en el modelo final no está la variable Provincia (significativa en el inicial), pero sí la Territorial (no era significativa en el inicial), esto puede darse debido a la alta relación que existe entre ellas (que es alta pero no tanto como para excluir alguna de las variables desde el inicio cuando se realizó el análisis descriptivo).

Ahora, comparando los modelos a través de los valores del AIC y pseudo-R^2, se tiene que:

	MODELO COMPLETO	MODELO BORUTA	MODELO REDUCCIÓN AIC
AIC	9859,7	9857,4	9811,67
Null deviance	10263,1	10263,1	10260
Residual deviance	9541,7	9545,4	9690
Pseudo-R^2	7,03%	6,99%	5,56%

Resultado 27:Comparación de las medidas de ajuste

Se puede observar que el modelo obtenido por extracción AIC, genera un menor valor de esta medida, y aunque su valor de Pseudo-R^2 es la más baja, hay que recordar que para este tipo de modelos no es la medida que mejor represente la capacidad predictiva del modelo.

Con base en los resultados Anova y los valores AIC, se **selecciona el modelo obtenido por el método de extracción AIC como el modelo óptimo de predicción**. Esta selección se basa en que la mayor parte de las variables que deja para estimar son identificadas como significativas (tanto en el modelo inicial (con todas las variables) como en el modelo Boruta), es el modelo con menor AIC y solo reduce en un 1% el pseudo-R^2 (versus los otros dos modelos), pero logrando una reducción importante en las variables (elimina 9 de las 19 variables totales).

7. RESULTADOS DEL MODELO SELECCIONADO SOBRE EL CONJUNTO DE ENTRENAMIENTO

Para la definición del punto de corte de la probabilidad, cuando se tiene un conjunto de datos homogéneo lo que se suele definir es que, si un caso estudiado supera el 50%, se clasifica en la categoría de "éxito", sin embargo, cuando se trata de conjuntos de datos asimétricos esto no se debe hacer ya que las estimaciones siempre van a estar sesgadas a la categoría mayoritaria.

Una forma sencilla de asignar el valor de la probabilidad para clasificar un caso como "éxito", en este caso como Fraude, es utilizar el porcentaje clasificado como tal en el total de datos que se analizan, que en este caso en particular sería del 4.7%. Sin embargo, en este trabajo se propone una forma más técnica para seleccionar ese valor de probabilidad, ésta consiste en un análisis de costos asignados a los errores de clasificación del modelo. Para esta propuesta se utiliza un caso (llamado "caso base"), con costos simbólicos para facilitar el entendimiento de la propuesta.

7.1. Caso Base

Para definir los costos en los que incurriría la entidad al utilizar este modelo de predicción, se establece un caso base con los siguientes valores[10]:

- Costos medio por investigar un posible fraude: 200€
- Costo medio de un fraude en el ramo hogar: 1000€

La forma de asignar estos costos al modelo es multiplicando cada costo por el número de pólizas mal clasificadas (errores de clasificación del modelo), esto es:

- Cuando el modelo estima un fraude cuando en realidad no lo es (FP o Error tipo I), la empresa incurriría en un costo de investigar un caso que en realidad no será fraudulento.

- Cuando el modelo estima un caso como No Fraude, pero en realdad sí lo es (FN o Error tipo II), la empresa incurriría en un costo por falta de detección temprana. Este costo sería igual al valor del fraude, que en este caso sería de 1.000€, multiplicado por el total de casos que clasificó como falsos negativos.

[10] Estos valores no corresponden a datos reales, se asignan tratando de guardar una proporción coherente (10 a 2 en este caso), de lo que resulta más costoso para la compañía que es la realización de un fraude.

Para explicar mejor cómo funciona el modelo aplicado a este caso base, se realiza el ejercicio analizando lo que le pasa a una póliza cuando el modelo la clasifica en cada uno de los escenarios posibles, esto es:

Coste medio por investigación:	200 €
Coste medio por Fraude Realizado	1.000 €

ESCENARIO	Costo investigación	Costo fraude	Pérdida / Ganancia
Verdadero - Positivo	-200	1000	800
Verdadero - Negativo	0	0	0
Falso - Positivo	-200	0	-200
Falso - Negativo	0	-1000	-1000

Tabla 22: Efecto económico de una póliza clasificada cada escenario posible

- **Escenario Verdadero Positivo**: aquí el modelo clasifica la póliza como fraude y en realidad sí es un fraude, por lo tanto, la empresa tendrá que pagar 200€ por costo medio de investigar el caso, pero como es un fraude que evita, está dejando de pagar 1.000€. Calculando lo que invierte (200€) menos lo que no paga por fraude (1.000€), la empresa está evitando pagar 800€ en total.

- **Escenario Verdadero Negativo**: en este escenario la empresa no tiene ni pérdida ni ganancia, ya que como no va a haber ninguna investigación ni se va a presentar un fraude, el valor neto es cero.

- **Escenario Falso Positivo:** aquí el modelo realizó una mala clasificación indicándole a la compañía que debía investigar este caso cuando en realidad no representaba ningún riesgo de fraude. El costo asociado a este error se corresponde con el coste medio de investigación de 200€.

- **Escenario Falso Negativo:** aquí el modelo comete un error al clasificar un Fraude como si no lo fuera. Este error (en función de los costos asumidos) es más grave que el anterior ya que resulta más costoso, teniendo una pérdida por fraude realizado de 1.000€.

7.2. Matriz de confusión y métricas asociadas.

Teniendo definidos los costos, se procede con la implementación del código de optimización, con el que se pretende minimizar la pérdida económica, para la definición del valor de la probabilidad. Este proceso consiste en un ciclo en el que se van evaluando los diferentes valores de la probabilidad hallada (la cual va desde el valor mínimo al máximo del vector de valores predicho, con un paso de 0,01[11]), y para cada valor calcula la matriz de confusión y multiplica los errores de clasificación por los costos determinados. El ciclo termina cuando evalúa el valor máximo de la probabilidad del vector de valores predicho y genera como resultado:

- El valor de la probabilidad que minimiza la pérdida,
- El valor monetario de la pérdida (que se tendría con dicha probabilidad),
- Cuadrantes de matriz de confusión.

En este caso, en el que se está aplicando el modelo a los datos de entrenamiento, se obtiene lo siguiente:

- **Valor de la probabilidad**: 0.18 (valor sobre el cual una póliza se considera fraude)

- **Matriz de confusión:**

		ESTIMADO		
		Fraude	No Fraude	Total Real
REAL	Fraude	74	1.203	1.277
	No Fraude	250	25.241	25.491
	Total Estimación	324	26.444	26.768

Resultado 28: Matriz de confusión - Conjunto de entrenamiento

Medidas de la matriz:

Exactitud	94,57%
Tasa de error	5,43%
Sensibilidad	5,79%
Especificidad	99,02%
Precisión	22,84%
Valor de predicción negativo	95,45%

Esta tabla de resultados indica que el modelo clasifica correctamente el 94.57% de las observaciones, sin embargo, en esta medida de exactitud se da el mismo peso a

[11] Este valor es parametrizable y depende del nivel de precisión que se quiera tener en el valor de la probabilidad.

las clasificaciones correctas de ambas categorías de la variable respuesta, es por esto que se recomienda analizar el porcentaje de las clasificaciones correctas en cada categoría por separado, esto es: el modelo clasifica bien las pólizas que efectivamente son fraude (verdaderos positivos, que se mide con la medida de Sensibilidad) en un 5.79%, y las que no lo son en un 99.02%. Como se puede observar, bajo este criterio de asignación de probabilidad por costos, se obtiene un modelo en el que los resultados de las predicciones para la categoría No Fraude son mucho más acertadas (esto se da por el sesgo que hay de los datos en esta categoría).

En cuanto a la tasa de error del modelo, (lo que el modelo clasifica mal) se puede ver que es relativamente baja (5.43%), sin embargo, esta afirmación solo aplica como conclusión basada en los datos porcentuales obtenidos con el modelo. Es decir, si se analiza desde una perspectiva de costos, este porcentaje es relativo ya que puede considerarse alto o bajo dependiendo de la empresa y los costos que tengan para asignar a cada error de estimación.

Curva ROC: Es otra forma de analizar los resultados del modelo pero de una forma visual, se basa en un análisis de la medida de Sensibilidad (frecuencia relativa de acertar la ocurrencia de Fraude- VP), comparada con la medida de [1- Especificidad] (frecuencia relativa de No Acertar la ocurrencia - FP). Obtener una curva ROC igual a la diagonal es asumir que se tiene la misma probabilidad de ocurrencia o no (es como tirar una moneda al aire), cualquier valor que esté por encima de la diagonal es una mejora, donde el mejor modelo posible estará situado en el punto con coordenadas (0,1).

La curva ROC del modelo obtenido es la siguiente:

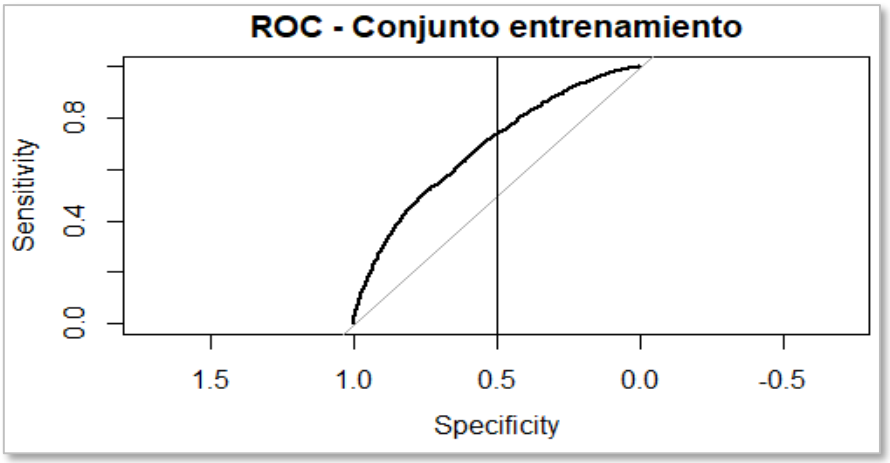


Imagen 24: Curva ROC - Conjunto de entrenamiento

Gráficamente se puede decir que es un buen modelo, pero para saber qué tan "bueno" es, se calcula el área bajo la curva (AUC), éste es un parámetro que sirve para evaluar la bondad de ajuste, en donde 1 significa que es una prueba perfecta y 0.5 inútil.

El AUC de este modelo es **0.68**. Si debajo de la curva se tiene el 0.5, el valor de 0.68 significa que este modelo predice un 0.18 mejor que el azar, es decir, **la capacidad predictiva de este modelo es 18% más alta.** Este porcentaje puede parecer poco, sin embargo, al tener datos asimétricos es normal que esto ocurra (Charpentier, 2019), y podría considerarse como aceptable para determinar que es un buen modelo.

7.3. Efecto Monetario por uso del modelo

Con los datos obtenidos en la matriz de confusión y los costos por errores de clasificación, se puede realizar el análisis de la conveniencia o no de aplicar este modelo en la compañía (en función de la pérdida o ganancia que pueda generar).

La pérdida que se genera por los errores de clasificación del modelo es -**1.253.000€**, la cual se calcula como:

Pérdida por utilizar el modelo

$$= Costo\ por\ investigar\ un\ caso\ que\ no\ va\ a\ ser\ fraude + costo\ por\ fraude\ no\ estimado$$

Esto es:

$$(250 * -200€) + (1.203 * -1000€) = -1.253.000€$$

A esta pérdida hay que sumarle lo que el modelo estimo bien y le ahorra a la compañía, esto es, casos que el modelo estimó como fraude y que en realidad lo eran. Como estos casos tendrán que ser investigados, aunque hayan sido bien clasificados, el ahorro sería

Ahorro por utilizar el modelo

$$= (pólizas\ predichas\ como\ fraude\ que\ sí\ lo\ son * costo\ de\ investigacion) - (pólizas\ predichas\ como\ fraude\ que\ sí\ lo\ son * costo\ del\ fraude\ evitado)$$

Esto es:

$$(74 * -200€) - (74 * -1000€) = -14.800 + (74.000) = 59.200€$$

En total, el costo en el que incurriría la entidad por utilizar el modelo sería:

$$-1.253.000 + 59.200 = -1.193.800€$$

Si se compara esta pérdida por errores del modelo (1.193.800€), con la pérdida de no tener el modelo (1.277 * 1000€ = 1.277.000€), se puede observar que, si la empresa decide utilizar una herramienta de predicción del fraude como ésta, estaría obteniendo un beneficio por usar el modelo que, en este caso académico, sería de 83.200€ (dinero que deja de perder por una buena clasificación de las pólizas), que corresponde a un ahorro del 6.52%.

Con base en los resultados obtenidos de la matriz de confusión, sus métricas, la curva ROC y el efecto monetario, se concluye que es un buen modelo y se procede con su aplicación en el conjunto de validación.

8. VALIDACIÓN DEL MODELO SELECCIONADO EN EL CONJUNTO DE TEST.

Para saber si el modelo es una buena herramienta de clasificación se evalúa en el conjunto de validación, para esto se realiza la predicción de la probabilidad al conjunto de datos y, utilizando el valor de probabilidad hallado (0,18), se construye la matriz de confusión, sus métricas, la curva ROC, el AUC y se calcula el efecto monetario en el que se incurriría por los errores de clasificación.

8.1. Matriz de confusión y métricas asociadas.

Aplicando el modelo obtenido en el conjunto de entrenamiento sobre el conjunto de validación se obtiene el siguiente resultado:

- **Valor de la probabilidad**: 0.18 (valor sobre el cual una póliza se considera fraude hallado en el entrenamiento del modelo)

- **Matriz de confusión:**

		ESTIMADO		
		Fraude	No Fraude	Total Real
REAL	Fraude	22	499	521
	No Fraude	114	10.837	10.951
	Total Estimación	136	11.336	11.472

Resultado 29: Matriz de confusión - Conjunto de validación

Medidas de la matriz

Exactitud	94,66%
Tasa de error	5,34%
Sensibilidad	4,22%
Especificidad	98,96%
Precisión	16,18%
Valor de predicción negativo	95,60%

Esta tabla de resultados indica que el modelo clasifica correctamente el 94.66% (considerado un buen porcentaje de clasificación en términos generales), pero, como se mencionó en el análisis realizado sobre el conjunto de entrenamiento, se recomienda analizar también el porcentaje de las clasificaciones correctas en cada categoría por separado, esto es: el porcentaje de casos que son fraude bien clasificados (Sensibilidad), el cual corresponde a un 4,22% y el porcentaje de casos

que no son fraude y quedaron bien clasificados (Especificidad), que corresponde a 98,96%. Se observa que el modelo predice muy bien cuando se tiene en cuenta solamente la medida de Exactitud, pero en cuanto a la predicción de la categoría de interés (Sensibilidad), es muy poco lo que puede estimar. Esto se debe a que el modelo está sesgando las predicciones a la categoría mayoritaria (No Fraude), que en este caso es en donde mejor realiza la estimación (Especificidad).

En cuanto a la tasa de error del modelo, se observa que el valor obtenido es 5.34%, que se podría considerar una medida baja de mala clasificación (si se le da el mismo peso a los FP y FN)

Curva ROC: Para este modelo se obtiene la siguiente curva:

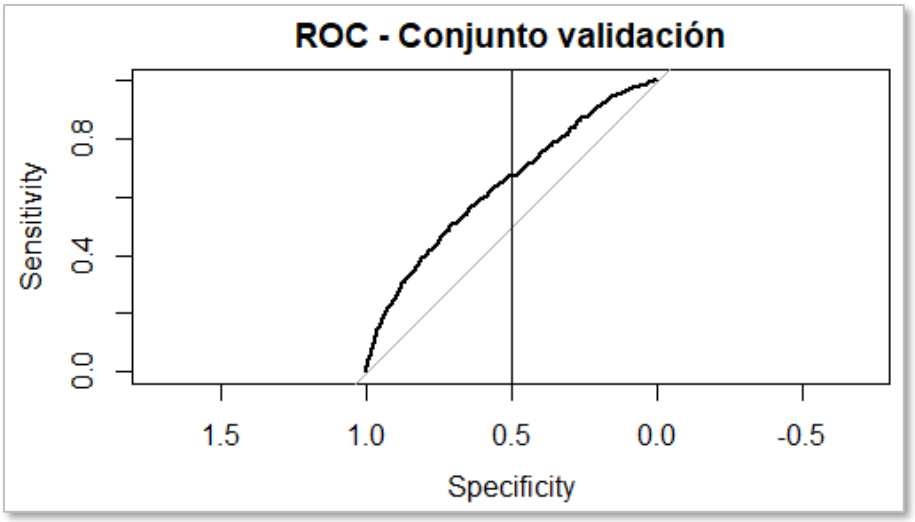


Imagen 25:Curva ROC - Modelo de validación

El AUC es **0.64**, si debajo de la curva se tiene el 0.5, el valor de 0.64 significa que este modelo predice un 0.14 mejor que el azar, es decir, **la capacidad predictiva de este modelo es un 14% más alta.**

8.2. Efecto Monetario por uso del modelo

La pérdida que se genera por los errores de clasificación del modelo es **-521.800€**, la cual se calcula como:

Pérdida por utilizar el modelo

$$= Costos\ por\ investigar\ un\ caso\ que\ no\ va\ a\ ser\ fraude + costos\ por\ fraude\ no\ estimado$$

Esto es:

$$(114 * -200€) + (449 * -1000€) = -521.800€$$

Ahora, si a este valor se le suma el ahorro que genera el modelo por clasificar adecuadamente los VP, se tiene que:

$$\textbf{\textit{Ahorro por utilizar el modelo}} = (\textit{pólizas predichas como fraude que sí lo son} * \textit{costo de investigacion}) - (\textit{pólizas predichas como fraude que sí lo son} * \textit{costo del fraude evitado})$$

Esto es:

$$(22 * -200€) - (22 * -1000€) = -4.400 + (22.000) = 17.600€$$

En total, el costo en los que incurriría la entidad por utilizar el modelo sería:

$$-521.800 + 17.600 = -504.200€$$

Si se compara esta pérdida por errores del modelo (511.400€), con la pérdida de no tener el modelo (521 * 1000€=521.000€), se puede observar que, si la empresa decide utilizar una herramienta de predicción del fraude como ésta, estaría obteniendo un beneficio por usar el modelo que, en este caso académico, sería de 16.800€ (dinero que deja de perder por una buena clasificación de las pólizas), que corresponde a un ahorro del 3.22%.

8.3. Comparación de resultados: conjunto de entrenamiento vs. conjunto de validación.

Ahora, para concluir si el modelo realiza una buena clasificación, se muestra la siguiente tabla que resume los resultados obtenidos en cada conjunto:

Métrica	Entrenamiento	Validación
Exactitud	94,57%	94,66%
Tasa de error	5,43%	5,34%
Sensibilidad	5,79%	4,22%
Especificidad	99,02%	98,96%
Precisión	22,84%	16,18%
Valor de prediccion negativo	95,45%	95,60%
Ahorro Monetario	6,52%	3,22%

Resultado 30:Métricas de clasificación del modelo en ambos conjuntos analizados.

Como se puede observar, el modelo conserva los buenos resultados de clasificación cuando se implementa sobre el conjunto de validación, especialmente en las medidas de Exactitud, Especificidad y Valor de predicción negativo, es por esto que se define como el modelo óptimo para la predicción del fraude en pólizas nuevas del ramo hogar.

En cuanto al porcentaje de ahorro monetario, no es pertinente dar conclusiones frente a si es mejor o peor en función del valor porcentual ya que está construido sobre un caso base en donde los costos asignados son solamente para efectos académicos. Lo que sí se puede concluir es que, independiente del porcentaje, implementar este modelo sí genera un ahorro económico para la compañía.

9. FUNCIÓN LOGIT OBTENIDA CON EL MODELO PARA LA CLASIFICACIÓN DE NUEVAS PÓLIZAS. DEFINICIÓN Y EJEMPLOS.

La función logit obtenida con el modelo es la siguiente[12]:

$$\eta = \log\left(\frac{\pi}{1-\pi}\right) = -5{,}5701 + \beta_i * Antigüedad_i + \beta_j * Continente_j + \beta_k * Tipo\_Mediador_k + \beta_l * Forma\_pago_l + \beta_m * Perfil_m + \beta_n * Territorial_n + \beta_p * Ubicación_p + \beta_q * Tipo\_uso_q + \beta_r * Edad_r + \beta_s * Valor\_cliente_s$$

Con:

- $i = \{1,2,\dots,11\}$
- $j = \{1,2,\dots,15\}$
- $k = \{Agentes, Axa\ Exclusive, Directo, Empleados\}$
- $l = \{2,3,4,6\}$
- $m = \{CPM, CVD, ERR, General, PM, Sinper, SVD, VD\}$
- $n = \{1,2,\dots,7\}$
- $p = \{1,2,3\}$
- $q = \{Habitual, Otro, Secundaria\}$
- $r = \{0,3,4,5,6,7\}$
- $s = \{A, B, C, D, E, F, G, X, Z\}$

Y, a partir de esta, se calcula la probabilidad π (probabilidad de ser fraude), tal que:

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

Tomando como referencia el valor de la probabilidad hallado por el método de optimización de costos en el modelo de entrenamiento (0.18), se tiene que para una póliza con valor $\pi \geq 0.18$, esa póliza será clasificada como “Fraude”.

A modo de ejemplo, se presentan 4 casos en los que el modelo clasifica la póliza de forma correcta, 2 de ellos en Fraude y los otros 2 en No fraude:

[12] El valor de los coeficientes beta se encuentra en el anexo 14.3.

Póliza 1: **Fraude**

Variables	Categoría	Descripción	Coeficientes
Intercepto			-5,57016255
antiguedad_vivienda	8	35_ 40 años	0,58068359
Continente_Cod	15	>250000	1,81269287
Tipo_Mediador	AGENTES	AGENTES	0
forma_pago	4	Trimestral	0,82508233
Perfil06	CPM	CPM	0
territorial	5	Sur	0,48195967
ubicación	1	Casco urbano	0
Tipo_Uso	Habitual	Habitual	0
Edad_Tomador	4	[26,50) años	0,52273368
CVM	A	A	0

Fn. Logit [η=log(π/(1-π))]
-1,347010425
Prob [π= e^η/(1+e^η)]
0,206359561
Prob Ref = 0,18
¿Fraude?
0,18778 > 0,18 → **SI**

Resultado 31: Ejemplo de predicción. Póliza clasificada como Fraude. Valor del cliente Alto

Póliza 2: **Fraude**

Variables	Categoría	Descripción	Coeficientes
Intercepto			-5,57016255
antiguedad_vivienda	8	35_ 40 años	0,58068359
Continente_Cod	13	(220000_240000]	1,68506937
Tipo_Mediador	AGENTES	AGENTES	0
forma_pago	3	Semestral	0,27130479
Perfil06	PM	PM	-0,03125256
territorial	7	Oeste	0,3231098
ubicación	1	Casco urbano	0
Tipo_Uso	Habitual	Habitual	0
Edad_Tomador	6	[60,70) años	0,22706214
CVM	F	F	1,34700069

Fn. Logit [η=log(π/(1-π))]
-1,167184725
Prob [π= e^η/(1+e^η)]
0,237364235
Prob Ref = 0,18
¿Fraude?
0,2266 > 0,184 → **SI**

Resultado 32: Ejemplo de predicción. Póliza clasificada como Fraude. Valor del cliente Bajo.

Póliza 3: **No Fraude**

Variables	Categoría	Descripción	Coeficientes
Intercepto			-5,57016255
antiguedad_vivienda	3	10_ 15 años	0,26240848
Continente_Cod	9	(140000_160000]	1,38255514
Tipo_Mediador	AGENTES	AGENTES	0
forma_pago	2	Anual	0
Perfil06	General	General	0,1154318
territorial	3	Levante	0,32753354
ubicación	1	Casco urbano	0
Tipo_Uso	Habitual	Habitual	0
Edad_Tomador	4	[26,50) años	0,52273368
CVM	A	A	0

Fn. Logit [η=log(π/(1-π))]
-2,959499912

Prob [π= e^η/(1+e^η)]
0,049289435

Prob Ref = 0,18

¿Fraude?
0,01518 < 0,184 → NO

Resultado 33:Ejemplo de predicción. Póliza clasificada como No Fraude. Valor del cliente Alto

Póliza 4: **No Fraude**

Variables	Categoría	Descripción	Coeficientes
Intercepto			-5,57016255
antiguedad_vivienda	2	5_ 10 años	0,37414031
Continente_Cod	9	(140000_160000]	1,38255514
Tipo_Mediador	AGENTES	AGENTES	0
forma_pago	2	Anual	0
Perfil06	VD	VD	0,2691018
territorial	4	Centro	0,16100073
ubicación	1	Casco urbano	0
Tipo_Uso	Secundaria	Secundaria	-0,36959318
Edad_Tomador	7	>=70 años	0,15645318
CVM	F	F	1,34700069

Fn. Logit [η=log(π/(1-π))]
-2,249503877

Prob [π= e^η/(1+e^η)]
0,095392268

Prob Ref = 0,18

¿Fraude?
0,02356 < 0,184 → NO

Resultado 34:Ejemplo de predicción. Póliza clasificada como No Fraude. Valor del cliente Bajo.

10. ANÁLISIS GRÁFICO PARA DETERMINAR EL VALOR DE LA PROBABILIDAD.

Una forma que puede resultar más sencilla para definir el valor de la probabilidad a partir de la cual una póliza se considerará fraude, es utilizando un gráfico de histogramas [13].

Esta propuesta consiste en aplicar la función logit a todo el conjunto de datos y calcular para cada póliza la probabilidad de ser clasificada como fraude, luego se grafican los histogramas de cada categoría de la variable respuesta con el valor de probabilidad hallado y se observan los puntos de corte de ambos histogramas, a partir de los cuales se detallan rangos sobre los que se proponen acciones puntuales.

La gráfica obtenida, después de aplicar la función logit, es la siguiente:

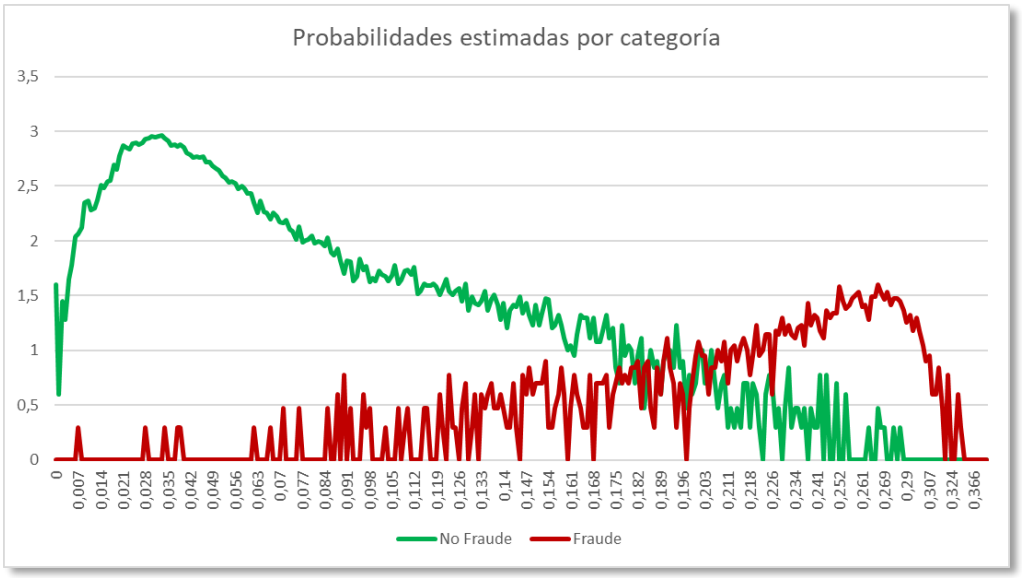


Imagen 26: Histogramas de probabilidades en el total del conjunto de datos.

En el eje *x* se encuentra el valor de las probabilidades estimadas hallado con la función logit y en el eje *y* el logaritmo del número de pólizas de cada valor de probabilidad[14]

[13] También se podrían utilizar los gráficos de boxplot (ver anexo 14.4), sin embargo, para este caso en donde el intervalo de la probabilidad estimada es tan corto, no son tan fáciles de identificar los cortes de los diferentes rangos.
[14] Se utiliza el logaritmo para reducir la escala y así hacer más visual el resultado.

Con base en esta gráfica se propone lo siguiente:

- **Probabilidad ≤ 0.06**: Son pólizas que no representan ninguna alerta de futuro Fraude, por lo que no se requeriría ningún tipo de acción.

- **Probabilidad en el intervalo (0.6,0.153**]: Es más probable que la póliza no sea fraude, aunque existe una menor probabilidad de que sí lo sea. En este intervalo no se requeriría ninguna acción, sin embargo, si la probabilidad está más cerca del límite superior, se recomienda hacer un análisis puntual de dicha póliza.

- **Probabilidad en el intervalo (0.153, 0.218**]: En este intervalo no es posible determinar si una póliza será o no fraude, es un rango en el que existe una probabilidad 50/50, por lo tanto, el analizar o no las pólizas de este rango serán una decisión que irá en función de los objetivos del análisis y las implicaciones de hacerlo o no.

- **Probabilidad > 0.218:** aunque hay una probabilidad de que la póliza no sea fraude, esta probabilidad es muy pequeña comparada con la de que sí sea fraude, por lo que se recomienda analizar todas las pólizas que superen este valor.

Con los argumentos anteriores, si se requiriera proponer un valor de la probabilidad éste sería el **0,218** (la cual es superior a la propuesta con el análisis de optimización de costos que es de 0,18). Sin embargo, determinar un único valor no es el objetivo de este análisis gráfico.

Si se pretende utilizar esta técnica para definir el valor de la probabilidad, se recomienda tener muy claros los objetivos que se tengan con el análisis (económicos, comerciales, estrategias de marketing, etc.), y las consecuencias de definir uno u otro valor (en costos, reputación, etc.), ya que, dependiendo de esto, el valor puede ser menor o mayor al que se recomienda en este trabajo.

11. CONCLUSIONES

De este trabajo se pueden extraer las siguientes conclusiones:

- Por lo general, los análisis de fraude se realizan a partir de la ocurrencia de un siniestro ya que esto genera, para el tomador, una oportunidad más clara de realizar el fraude y, para la entidad, una predicción de posible ocurrencia con mayor asertividad, sin embargo, con este trabajo se demuestra que es posible realizar una buena estimación del fraude desde la suscripción de la póliza, utilizando, principalmente, variables propias del objeto asegurado, en este caso la vivienda.
- Implementar una herramienta de detección de fraude desde el momento de la suscripción ayuda a la compañía a disminuir costos, ya sea evitando los fraudes por una investigación a tiempo, o tomando decisiones más radicales como trasladar los costos al tomador de la póliza desde el inicio (incremento en la tarifa) o decidir no realizar el contrato de seguro con dicho tomador.
- El modelo de regresión logística desarrollado en este trabajo puede considerarse como una buena herramienta de clasificación ya que genera muy buenos resultados en su globalidad. Es de resaltar que la medida de sensibilidad, que es la predicción de la categoría de Fraude propiamente, no genera muy buenos resultados, sin embargo, por la característica de asimetría de los datos, la conclusión de que éste es un buen modelo se hace teniendo en cuenta todas las métricas en su conjunto y no solo esta medida en particular.
- Determinar el valor de la probabilidad en función de los costos asociados a los errores de clasificación del modelo es una buena estrategia cuando se decide realizar acciones puntuales que implican costos monetarios. Esto se debe a que, con la optimización, se logra disminuir la posible pérdida en las que se incurriría por utilizar esta herramienta.
- Realizar análisis gráficos para determinar el valor de la probabilidad, ya sean histogramas, boxplot u otro tipo de gráficos, puede resultar más sencillo de implementar, sin embargo, esta forma de determinar la probabilidad no se recomienda cuando hay implícitos costos elevados y no se ha realizado un análisis previo de los mismos.

En definitiva, este tipo de herramientas de predicción del fraude desde la suscripción de la póliza, y no solo desde la ocurrencia de un siniestro, deberían incluirse como un input más en los procesos de comercialización y tarificación de la compañía debido a los grandes beneficios que genera (principalmente de tipo económico).

12. FUTUROS TRABAJOS

A partir de este trabajo se pueden realizar otro tipo de análisis que podrían ayudar a mejorar los resultados obtenidos, por ejemplo:

- Incorporar variables adicionales del tomador de la póliza:

    - Propias del tomador: ingresos periódicos, estado civil, ocupación.
    - Relación del tomador con la compañía: si es un cliente que ya tiene otros productos con la compañía sería interesante tener dentro del análisis variables como: antigüedad del cliente, número de ramos contratados, número de siniestros asociados al cliente, número de siniestros investigados por posible fraude, número de fraudes identificados por cliente.

- Hacer un análisis condicional de la probabilidad de fraude desde la suscripción, incluyendo la probabilidad de ocurrencia del siniestro (un modelo intermedio que genere la probabilidad de siniestro y posterior fraude)

- En la optimización por costos, considerar una penalización por tipo de error tal que se castigue el error de clasificación que más costo genere para la entidad que utiliza el modelo.

13. BIBLIOGRAFÍA

Álvarez Jareño, J. A. (2019). Aplicación de métodos estadísticos, económicos y de aprendizaje automático para la detección del fraude. *Aprendizaje Automático aplicado al Sector Asegurador. Ciclo de primavera de conferencias del MCAF.* Universidad Complutense de Madrid, Madrid.

Álvarez, B. (2018). *Engañar a tu seguro tiene un precio. Consumer. Economía y Consumo.* Obtenido de https://www.consumer.es/economia-domestica/servicios-y-hogar/enganar-a-tu-seguro-tiene-un-precio.html

AXA España. (2018). *V Mapa AXA del Fraude en España.* Obtenido de https://www.axa.es/documents/1119421/143282252/V+Mapa++AXA+del+Fraude+en+Espa%C3%B1a.pdf

AXA España. (2019). *VI Mapa AXA del Fraude en España.* Obtenido de https://www.axa.es/documents/1119421/160164507/VI+Mapa+AXA+del+Fraude+en+Espa%C3%B1a_informe.pdf/497b0e40-bb30-f9ab-1d52-b188adb0c391

Ayuso Gutiérrez, M. (1998). *Modelos econométricos para la detección del fraude en el seguro del automóvil.* Universidad de Barcelona, Barcelona.

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

Brownlee, J. (2015). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. Machine Learning Mastery*. Obtenido de https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Cañadas Reche, J. L. (2013). *Regresión logística. Tratamiento computacional con R.* (Trabajo fin de Máster), Universidad de Granada, Granada.

Charpentier, A. (2019). *On the poor performance of classifiers in insurance models.* Obtenido de https://freakonometrics.hypotheses.org/57954

Daymon, J. (2018). Diferencias entre el aprendizaje supervisado y el aprendizaje no supervisado. *Diferencia Entre*. Obtenido de http://www.diferenciaentre.net/diferencias-entre-el-aprendizaje-supervisado-y-el-aprendizaje-no-supervisado/

De la Espriella, C. (2012). *Fraude en seguros. Una aproximación al caso colombiano.* FASECOLDA.

DEJ. (2019). *Fraude. Diccionario del español jurídico de la real academia española*. Obtenido de https://dej.rae.es/lema/fraude

Departamento de Economía. (s.f.). *Economía Aplicada. Modelos con variables dependientes binarias. Universidad Carlos III de Madrid*. Recuperado el 20 de 08 de 2019, de

http://www.eco.uc3m.es/docencia/EconomiaAplicada/materiales/ModelosProbabilidad.pdf

Developers Google. (s.f.). *Clasificación ROC y AUC*. Recuperado el 15 de 08 de 2019, de https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419

Díaz Sanjuán, A. (2018). *Modelos de predicción para la detección y estudio del fraude en el seguro de automóviles.* (Trabajo fin de Máster), Universidad de Valencia, Valencia.

Díaz, Z. (2019). Modelos lineales generalizados para tarificación en seguros. *Apuntes de la asignatura: Aplicaciones Actutariales (MCAF).* Universidad Complutense de Madrid.

Dobson, A. J. (2001). *An introduction to generalized lineal models* (2da ed.). Boca Raton : Chapman & Hall/CRC.

Equipo de redacción MV. (s.f.). *Fraudes más comunes a las compañías de seguros. MV Aseguradores*. Recuperado el 16 de 08 de 2019, de https://www.mvaseguradores.com/fraudes-companias-seguros/#Fraudes_en_los_seguros_de_hogar

García, A. M. (2017). *Técnicas estadísticas para la detección del fraude.* (Trabajo Fin de Máster), Universidad Complutense de Madrid., Facultad de estudios estadísticos, Madrid.

Grimaldi, A. (2019). *Axa cifra el fraude asegurador en Andalucía en 11,6 millones en 2018. Diario de Sevilla.* Obtenido de https://www.diariodesevilla.es/economia/Axa-fraude-asegurador-Andalucia-2018_0_1356764905.html

Guerrero, J. A. (2016). *El problema de la dimensionalidad. Índice. Revista de estadística y sociedad.* Obtenido de http://www.revistaindice.com/numero68/p22.pdf

Heras, A. (2019). Regresión logística. *Apuntes de la asignatura: Aplicaciones Actutariales (MCAF).* Universidad Complutense de Madrid, Madrid.

Interbrand. (2018). *Interbrand lanza el informe Best Global Brands 2018*. Obtenido de https://www.interbrand.com/es/newsroom/interbrand-lanza-el-informe-best-global-brands-2018/

Kuhn , M., Chow , F., & Wickham, H. (2019). *rsample: General Resampling Infrastructure.* R package version 0.0.5. Obtenido de https://CRAN.R-project.org/package=rsample

Martínez de la Puente, F. (s.f.). *El fraude en los seguros. Seguros CEA*. Recuperado el 02 de 09 de 2019, de https://www.seguroscea.es/blog/280-el-fraude-en-los-seguros

Miron B. , K., & Witold R. , R. (2010). Feature Selection with the {Boruta} Package. *Journal of Statistical Software, 36*(11), 1-13. Obtenido de http://www.jstatsoft.org/v36/i11/

Morales, J. (2018). *GLM para respuesta binaria. Modelos estocásticos. Grado biotecnología*. Obtenido de https://rstudio-pubs-static.s3.amazonaws.com/366804_ad2ed31b93184115a0a44a3a8b10750e.html

Pathak, M. (2018). *Feature Selection in R with the Boruta R Package. DataCamp.* Obtenido de https://www.datacamp.com/community/tutorials/feature-selection-R-boruta#comments

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Obtenido de https://www.R-project.org/

RStudio Team. (2018). *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA. Obtenido de http://www.rstudio.com/

Santacruz, A. (2016). *Por qué es importante trabajar con datos balanceados para clasificación. AMSANTAC.* Obtenido de http://amsantac.co/blog/es/2016/09/20/balanced-image-classification-r-es.html

Sector Asegurador. (2015). *El fraude en los seguros al detalle.* Obtenido de https://www.sectoraseguradores.es/fraude-en-seguros/

Soria, A. (2019). *Fraude en seguros de coche.* Obtenido de https://www.arpem.com/noticias/seguros/sabias-que/fraude-seguros-coche-3390694-n.html

UNESPA. (2019). *Asociación Empresarial del Seguro. XXV Concurso de detección de fraudes de ICEA.* Obtenido de http://www.unespa.es/notasdeprensa/xxv-premios-deteccion-fraudes-seguro/

Wikipedia. (s.f.). *Criterio de información de Akaike. Wikipedia. Enciclopedia libre.* Recuperado el 21 de 08 de 2019, de https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_de_Akaike

Wikipedia. (s.f.). *Modelo Lineal Generalizado. Wikipedia. Enciclopedia libre.* Recuperado el 15 de 08 de 2019, de https://es.wikipedia.org/wiki/Modelo_lineal_generalizado

14. ANEXOS

14.1. Matriz de algoritmos de *machine learning*.

La siguiente matriz ayuda a diferenciar los diferentes algoritmos de *machine learning* a partir del tipo de variables que se tienen en los modelos y el tipo de aprendizaje (conocimiento o asignación de patrones), que se tiene de los datos.

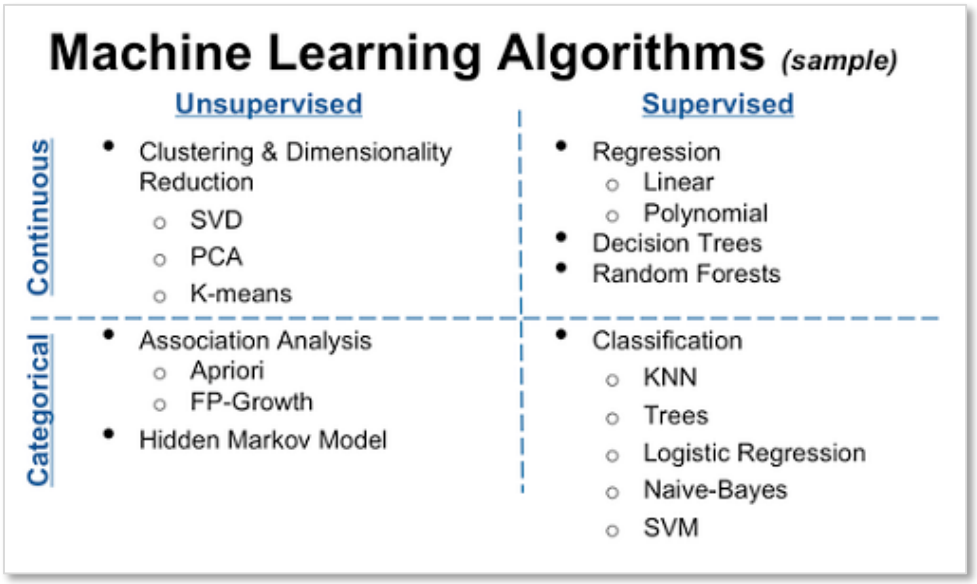


Imagen 27: Matriz de Algoritmos Machine Learning.
Fuente: *(Daymon, 2018)*

14.2. Resumen de la ejecución de optimización del modelo utilizando el método de extracción AIC.

	MODELO 1	MODELO 2	MODELO 3	MODELO 4	MODELO 5
	Fraude ~ antiguedad_vivienda + contenido + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + provincia + superficie + territorial + tipo_vivienda + ubicacion + Tipo_usuario +Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + contenido + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + provincia + territorial + tipo_vivienda + ubicacion + Tipo_usuario + Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + contenido + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + territorial + tipo_vivienda + ubicacion + Tipo_usuario + Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + territorial + tipo_vivienda + ubicacion + Tipo_usuario + Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + territorial + ubicacion + Tipo_usuario + Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM
AIC	**9859,72**	**9843,56**	**9832,73**	**9821,44**	**9818,7**
superficie	**9843.6**				
provincia	9632.9	**9832.7**			
contenido	9551.5	9833.2	**9821.4**		
territorial	9550.6	9840.2	9843.4	9832.8	9830.9
forma_pago	9545.4	9841.3	9831.7	9820.6	9818.7
Rehabilitada	9541.9	9841.8	9831.1	9820.0	9817.3
Tipo_usuario	9542.0	9841.7	9830.9	9819.5	**9816.8**
tipo_vivienda	9548.5	9842.1	9830.1	**9818.7**	
Hipoteca	9543.6	9843.3	9832.3	9820.9	9818.2
nota_bureau	9551.7	9843.6	9833.0	9821.6	9818.9
Perfil06	9556.2	9843.8	9836.8	9825.5	9822.3
nota_global	9561.6	9845.8	9833.1	9821.7	9818.2
antiguedad_vi	9563.9	9845.8	9834.9	9823.1	9819.8
Continente_C	9572.0	9855.3	9844.2	9835.4	9852.2
Tipo_Mediadc	9552.1	9848.0	9837.0	9825.7	9823.1
ubicacion	9554.2	9852.5	9840.2	9828.6	9822.9
Tipo_Uso	9556.5	9854.7	9844.4	9834.6	9830.1
Edad_Tomadc	9569.3	9860.8	9851.2	9839.9	9836.6
CVM	9856.5	0142.5	0133.9	0123.2	0119.2

	MODELO 6	MODELO 7	MODELO 8	MODELO 9	MODELO 10
	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + territorial + ubicacion + Tipo_Uso + Rehabilitada + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + nota_global + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + nota_bureau + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + Hipoteca + CVM	Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + CVM
AIC	**9816,8**	**9815,36**	**9814,88**	**9812,02**	**9811,67**
superficie					
provincia					
contenido					
territorial	9829.0	9827.6	9833.2	9830.5	9830.4
forma_pago	9816.8	9815.4	9815.4	9813.5	9813.4
Rehabilitada	**9815.4**				
Tipo_usuario					
tipo_vivienda					
Hipoteca	9816.3	9814.9	9814.6	**9811.7**	
nota_bureau	9817.0	9815.7	**9812.0**		
Perfil06	9820.4	9819.3	9818.9	9816.1	9815.7
nota_global	9816.3	**9814.9**			
antiguedad_v	9817.8	9817.9	9817.3	9814.3	9813.4
Continente_C	9869.8	9868.9	9866.1	9862.0	9863.5
Tipo_Mediadc	9821.2	9819.8	9819.4	9816.5	9816.1
ubicacion	9821.0	9819.5	9819.5	9816.6	9816.4
Tipo_Uso	9828.1	9826.8	9829.5	9826.5	9826.9
Edad_Tomadc	9834.6	9833.4	9833.7	9833.8	9837.3
CVM	0117.3	0116.0	0116.1	0115.8	0114.3

Resultado 35: Detalle del proceso de selección de variables utilizando el proceso de extracción AIC

El modelo final seleccionado es el Modelo 10 y es el utilizado para la validación en el grupo de entrenamiento. Este modelo corresponde a:

Modelo_10 = (Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + CVM)

14.3. Modelo final y tabla con los coeficientes resultantes.

El modelo obtenido como óptimo es el siguiente, el cual utiliza 10 de las 19 variables iniciales:

Fraude ~ antiguedad_vivienda + Continente_Cod + Tipo_Mediador + forma_pago + Perfil06 + territorial + ubicacion + Tipo_Uso + Edad_Tomador + CVM

Los coeficientes de cada variable y el intercepto, son los siguientes (estos son los que se utilizan para realizar las predicciones de Fraude en las futuras pólizas utilizando la función logit):

Variable	Categoría	Coeficientes
(Intercept)		-5,570162555
antiguedad_vivienda	1	0
antiguedad_vivienda	2	0,374140305
antiguedad_vivienda	3	0,262408483
antiguedad_vivienda	4	0,464164286
antiguedad_vivienda	5	0,208257267
antiguedad_vivienda	6	0,405086622
antiguedad_vivienda	7	0,578513452
antiguedad_vivienda	8	0,580683585
antiguedad_vivienda	9	0,39406281
antiguedad_vivienda	10	0,405051326
antiguedad_vivienda	11	-0,701820814
Continente_Cod	1	0
Continente_Cod	2	0,782421057
Continente_Cod	3	1,114267177
Continente_Cod	4	1,31782917
Continente_Cod	5	1,259719803
Continente_Cod	6	1,212804051
Continente_Cod	7	1,288238411
Continente_Cod	8	1,20911852
Continente_Cod	9	1,382555143
Continente_Cod	10	1,567256398
Continente_Cod	11	1,523428483
Continente_Cod	12	1,624350614

Continente_Cod	13	1,68506937
Continente_Cod	14	1,545407286
Continente_Cod	15	1,812692871
Tipo_Mediador	AGENTES	0
Tipo_Mediador	AXA EXCLUSIV	1,234819925
Tipo_Mediador	DIRECTO	0,395104806
Tipo_Mediador	EMPLEADOS	-1,468820852
forma_pago	2	0
forma_pago	3	0,271304793
forma_pago	4	0,825082326
forma_pago	6	-11,54552753
Perfil06	CPM	0
Perfil06	CVD	-0,015093481
Perfil06	ERR	-11,44999168
Perfil06	General	0,115431795
Perfil06	PM	-0,031252557
Perfil06	Sinperf	-0,319708893
Perfil06	SVD	0,296625546
Perfil06	VD	0,269101798
territorial	1	0
territorial	2	0,291899353
territorial	3	0,32753354
territorial	4	0,161000733
territorial	5	0,481959665
territorial	6	1,221982957
territorial	7	0,3231098
ubicación	1	0
ubicación	2	-0,341826731
ubicación	3	-11,33086538
Tipo_Uso	Habitual	0
Tipo_Uso	Otro	-0,08470519
Tipo_Uso	Secundaria	-0,369593175
Edad_Tomador	0	0
Edad_Tomador	3	0,62540825
Edad_Tomador	4	0,522733682
Edad_Tomador	5	0,128663524
Edad_Tomador	6	0,227062145
Edad_Tomador	7	0,156453181
CVM	A	0
CVM	B	-0,069244343
CVM	C	-0,115564061
CVM	D	0,402815544
CVM	E	-0,053836312
CVM	F	1,347000693
CVM	G	-1,077650632
CVM	X	-0,060151501
CVM	Z	0,377182646

14.4. Gráfico de Boxplot para determinar el valor de la probabilidad.

Este tipo de gráficos también podría utilizarse para determinar el valor de la probabilidad sobre la cual se podría determinar que una póliza será fraude, sin embargo, para este caso en particular, sería posible dar el valor a partir del cuartil 3 (0,75) del boxplot 1 (Fraude), pero no sería fácil establecer los demás rangos (como se hizo con los histogramas)

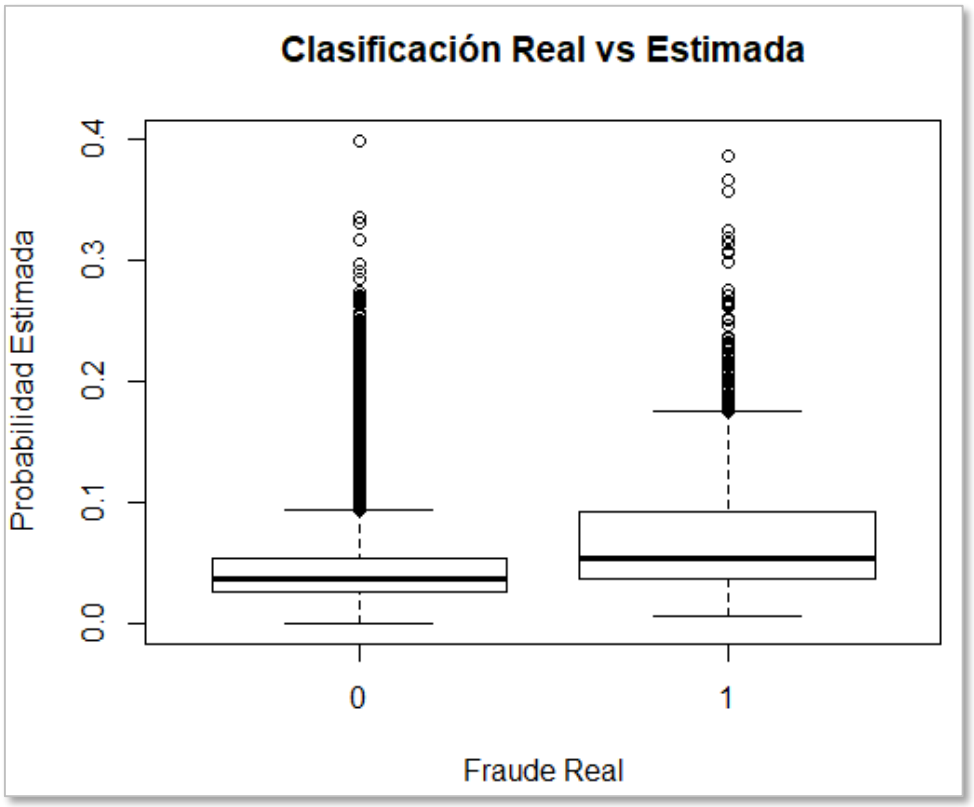


Imagen 28:Gráfico boxplot para determinar el valor de la probabilidad.

14.5. Código en R utilizado para la modelación.

Para la construcción de este código se utilizaron como referentes principales: (Álvarez Jareño, 2019; Díaz, 2019; Heras, 2019; Morales, 2018).

```
##################################################################################################
#################### Regresión Logística Con Datos Asimétricos: ##################################
##### Aplicación a la detección de Fraude al momento de la suscripción en seguros multi-riesgo ####
##################################### (Hogar).####################################################
################################## POR: Johana Osorno Gómez ######################################
##################################################################################################


#-------------------------------------------------------------------------
# CARGA DE LAS LIBRERIAS Y LOS DATOS
#-------------------------------------------------------------------------

options(scipen = 999)
set.seed(123)

# Clear workspace
rm(list = ls())

library(tidyverse)
library(rsample)
library(stats) a
library(rpart)
library(AUC)
library(pROC)
library(ROCR)
library(data.table)
library(funModeling)
library(ranger)
library(Boruta)
library(car)

indv= read.table(choose.files(),header=TRUE,sep=";")
names(indv)
dim(indv)

fix(indv)
#Con esta orden, se abre la tabla-editor la cual muestra los datos cargados.
#Se debe cerrar para continuar
attach(indv)
#detach(indv)


#-------------------------------------------------------------------------
# ANÁLISIS PREVIO DE LAS VARIABLES
#-------------------------------------------------------------------------

#Para identificar las variables categóricas como tal, se aplica la funcion as.factor
labels(indv[2, ])
convert_indv <- c(1:21)
indv[, convert_indv] <- data.frame(apply(indv[convert_indv], 2, as.factor))

#Distribución Variable Respuesta: FRAUDE
table(Fraude)
prop.table(table(Fraude))
barplot(table(Fraude),col=c("skyblue4","lightgoldenrod2"),main="Fraude", border= FALSE)
```

```
#Análisis descriptivo de las variables
#-------------------------------------------------------------------------------

#Como las variables son de tipo cualitativo, se utiliza el test chi cuadrado para ver la relacion
#que existe entre ellas y la variable respuesta:

#Fraude vs antigüedad de la vivienda
chisq.test(antiguedad_vivienda, Fraude)
table(antiguedad_vivienda, Fraude)
barplot(table(Fraude, antiguedad_vivienda),main="Fraude vs Antigüedad de la vivienda",
        col=c("skyblue4","lightgoldenrod2"),col.main="lightcyan4",border= FALSE)


#Fraude vs contenido
chisq.test(contenido, Fraude)
table(contenido, Fraude)
barplot(table(Fraude, contenido),main="Fraude vs contenido",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Continente
chisq.test(Continente_Cod, Fraude)
table(Continente_Cod, Fraude)
barplot(table(Fraude, Continente_Cod),main="Fraude vs Continente",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Tipo de mediador
chisq.test(Tipo_Mediador, Fraude)
table(Tipo_Mediador, Fraude)
barplot(table(Fraude, Tipo_Mediador),main="Fraude vs Tipo_Mediador",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)

#Fraude vs Forma de pago
chisq.test(forma_pago, Fraude)
table(forma_pago, Fraude)
barplot(table(Fraude, forma_pago),main="Fraude vs Forma de pago",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


# Fraude vs nota_bureau
chisq.test(nota_bureau, Fraude)
table(nota_bureau, Fraude)
barplot(table(Fraude, nota_bureau),main="Fraude vs Nota Bureau",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs nota global
chisq.test(nota_global, Fraude)
table(nota_global, Fraude)
barplot(table(Fraude, nota_global),main="Fraude vs Nota Global",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Perfil06
chisq.test(Perfil06, Fraude)
table(Perfil06, Fraude)
barplot(table(Fraude, Perfil06),main="Fraude vs Clasificación del Mediador",
        col=c("skyblue4","lightgoldenrod2"),col.main="lightcyan4",border= FALSE)


#Fraude vs Provincia
chisq.test(provincia, Fraude)
table(provincia, Fraude)
barplot(table(Fraude, provincia),main="Fraude vs Provincia",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)
```

```
#Fraude vs Superficie
chisq.test(superficie, Fraude)
table(superficie, Fraude)
barplot(table(Fraude, superficie),main="Fraude vs Superficie",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Territorial
chisq.test(territorial, Fraude)
table(territorial, Fraude)
barplot(table(Fraude, territorial),main="Fraude vs Territorial",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Tipo de vivienda
chisq.test(tipo_vivienda, Fraude)
table(tipo_vivienda, Fraude)
barplot(table(Fraude, tipo_vivienda),main="Fraude vs Tipo de vivienda",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Ubicación
chisq.test(ubicacion, Fraude)
table(ubicacion, Fraude)
barplot(table(Fraude, ubicacion),main="Fraude vs Ubicacion",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Tipo Usuario
chisq.test(Tipo_usuario, Fraude)
table(Tipo_usuario, Fraude)
barplot(table(Fraude, Tipo_usuario),main="Fraude vs Tipo de Usuario",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Tipo Uso
chisq.test(Tipo_Uso, Fraude)
table(Tipo_Uso, Fraude)
barplot(table(Fraude, Tipo_Uso),main="Fraude vs Tipo de uso",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Rehabilitada
chisq.test(Rehabilitada, Fraude)
table(Rehabilitada, Fraude)
barplot(table(Fraude, Rehabilitada),main="Fraude vs Rehabilitada",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Edad del tomador
chisq.test(Edad_Tomador, Fraude)
table(Edad_Tomador, Fraude)
barplot(table(Fraude, Edad_Tomador),main="Fraude vs Edad_Tomador",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Hipoteca
chisq.test(Hipoteca, Fraude)
table(Hipoteca, Fraude)
barplot(table(Fraude, Hipoteca),main="Fraude vs HIPOTECA",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#Fraude vs Valor del cliente
chisq.test(CVM, Fraude)
table(CVM, Fraude)
barplot(table(Fraude, CVM),main="Fraude vs Valor del cliente",col=c("skyblue4","lightgoldenrod2"),
        col.main="lightcyan4",border= FALSE)


#-------------------------------------------------------------------------------
#DIVISION DE LOS DATOS EN CONJUNTO DE ENTRENAMIENTO Y CONJUNTO DE VALIDACION
#-------------------------------------------------------------------------------


#Se convierte la variable respuesta en variable dicotomica para poder utilizar el modelo GLM
indv$Fraude<-as.numeric(indv$Fraude)
indv$Fraude<-ifelse(indv$Fraude==1,0,1)
table(indv$Fraude)
```

```
#lo primero es crear una división del fichero para tener los conjuntos de entrenamiento y
#el de validacion:
  #1) train (70%): utilizado para entrenar los algoritmos, comparar los modelos y selección
  #   del modelo óptimo
  #2) test  (30%): conjunto de datos sobre el cual se valida el modelo seleccionado (modelo óptimo)

# En este caso, como la variable respuesta tiene una distribución sesgada se utiliza un
# Muestreo Estratificado para conservar las proporciones de Fraude en ambos conjuntos.
# Para esto se utiliza la función initial_split del paquete rsample


#Muestreo estratificado con el paquete rsample
#-------------------------------------------------

set.seed(123) #semilla utilizada para la generacion de los numeros aleatorios
split_strat <- initial_split(indv, prop = 0.7, strata = "Fraude")
train  <- training(split_strat)
test   <- testing(split_strat)

#Se valida que la variable respuesta es consistente en los grupos de entrenamiento y validacion:

# Conjunto de Entrenamiento
table(train$Fraude) %>% prop.table()
  ##        No       Yes
  ##0.95229378 0.04770622

#Conjunto de validación
table(test$Fraude) %>% prop.table()
  ##         No        Yes
  ##0.95458508 0.04541492
#-------------------------------------------------
#SELECCION DE LAS VARIABLES CON BORUTA
#-------------------------------------------------

train=train[,-c(21)] #Se quita la primera variable que corresponde al código de la póliza
names(train)
boruta.train <- Boruta(Fraude~., data = train, doTrace = 2)
print(boruta.train)

# Garfico de la Importancia de las Variables
plot(boruta.train, xlab = "", xaxt = "n")
lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)
  boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
names(lz) <- colnames(boruta.train$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.7)

getSelectedAttributes(boruta.train, withTentative = T)
borutadf<-attStats(boruta.train)
print(borutadf)

#Para sacar las métricas del modelo BORUTA
M1_Boruta <- glm(Fraude ~ antiguedad_vivienda + contenido + Continente_Cod + Tipo_Mediador +
                  nota_bureau + nota_global + Perfil06 + provincia + superficie + territorial +
                  tipo_vivienda + ubicacion + Tipo_usuario + Tipo_Uso + Rehabilitada + Edad_Tomador +
                  Hipoteca + CVM,family = binomial(link = logit),data= train)
summary(M1_Boruta)
Anova(M1_Boruta)

#-------------------------------------------------
#MODELO GLM SOBRE CONJUNTO DE ENTRENAMIENTO
#-------------------------------------------------

M1_i1 <- glm(Fraude~ .,family = binomial(link = logit),data= train)

#como el glm se almacena en una lista, se utiliza la siguiente linea para mostrar los elementos
#disponibles en la variable logit
lapply(M1_i1,class)[1:3]

# Resumen del modelo de entrenamiento ajustado
summary(M1_i1)


#Bondad de ajuste del modelo
#--------------------------------------

#Para evaluar la capacidad explicativa del modelo se debe comparar la NULL DEVIANCE con el cuantil 0.95
#de una chi cuadrado con #grados de libertad igual a los correspondientes a la RESIDUAL DEVIANCE.
#Esto es, se debe realizar el análisis a través del pseudo-R^2: se calcula:
#(1-(Devianza Residual/devianza nula) y funciona como el R^2 de las regresiones lineales.
#En este caso se tiene que:
  #Deviance asociada al modelo ajsutado=Null deviance=10263
  #Deviance que no es capaz de explicar el modelo ajustado = Residual deviance=9541
#por lo que el pseudo-R^2 sería:

pseudoR_M1_i1 = 1-(9541.7/10263.1)
pseudoR_M1_i1
```

```
#0.07029065. Por ser un modelo tan imbalanceado es normal que este valor sea bajo,
#ya que está directamente relacionado con la precisión del modelo y la predicción para casos de exito
#en este caso esta completamente sesgada a la variable mayoritaria (casos de fracaso)

#Significatividad del modelo
Anova(M1_i1)


#-------------------------------------------------------------
#SELECCION DE LAS VARIABLES POR REDUCCION DE AIC
#-------------------------------------------------------------

MFinal_i1 <- step(M1_i1, test = "Chisq")
#El proceso de selección de efectos indica que se pueden eliminar 9 variables,
# reduciendo el modelo a 10 varibles predictoras

MFinal_i1
  #Degrees of Freedom: 26767 Total (i.e. Null);  26710
  #Null Deviance:10260
  #Residual Deviance: 9690
  #AIC: 9812


#Se calcula el Pseudo-R^2
pseudoR_MFinal_i1 = 1-(9690/10260 )
pseudoR_MFinal_i1
  #0.052918295

#Anova
Anova(MFinal_i1)


#Calculo de los odds ratio
#----------------------------
exp(cbind(OR = coef(MFinal_i1)))

#-------------------------------------------------------------
# PREDICCIÓN EN CONJUNTO DE ENTRENAMIENTO
#-------------------------------------------------------------

#MFinal_i1
predicted_value  <-0
predicted_value <- predict(MFinal_i1,type = "response",newdata= train)

#Se convierte el valor de predicciones en una tabla
predictor=as.table(predicted_value)

#Se exporta a excel el vector de predicciones
write.csv(predictor, file="Vector_prediccion_TRAIN.csv",row.names=FALSE)

#boxplot real vs predicho
box <- data.frame(predictor,train$Fraude)
head(box,13) #valido que la observacion 12 sea fraude (1 en real)

#Grafico boxplot
boxplot(box$Freq~box$train.Fraude, main="Clasificación Real vs Estimada",
        xlab="Fraude Real", ylab="Probabilidad Estimada")


#Análisis de los resultados predichos
min <- min(predicted_value)
min
max <-max(predicted_value)
max
median(predicted_value)
mean(predicted_value) #[1] 0.04770385: que corresponde a la proporcion de datos marcados
#como fraude en el conjunto de entrenamiento

names(train)
head( predicted_value)
table(train$Fraude)
str(train$Fraude)
```

```
# CURVA ROC - TRAIN
#-----------------------

# require(AUC)
#curva ROC (1)
train$prob_fraude <- predicted_value
g <- roc(train$Fraude ~ train$prob_fraude , data = train)
plot(g, main="ROC - Conjunto entrenamiento")
abline(v=(0.5))

#AUC
AUC <- auc(train$Fraude, predicted_value)
AUC[1] # 0.6818098


#-------------------------------------------------------------------------------
# MATRIZ DE CONFUSIÓN Y OPTIMIZACION DEL MODELO EN EL CONJUNTO DE ENTRENAMIENTO
#-------------------------------------------------------------------------------

#Construccion de la matriz de confusión a partir de la optimizacion del algoritmo
#para hallar el valor de la probabilidad que optimiza la pérdida por mala clasificación (minimiza)


# OPTIMIZACIÓN PARA HALLAR LA MATRIZ CON EL MINIMO COSTO - TRAIN
#-------------------------------------------------------------------------------


prueba <- list()
count <- 1

for (prob_fraude in seq(min,max,0.01)){

  predicted_class <- ifelse(predicted_value>prob_fraude , "Yes","No")
  performance_data<-data.frame(observed=train$Fraude,
                               predicted= predicted_class)

  #Construccion de la matriz de confusión

  #Totales casos reales y predichos para respuesta positiva (Fraude) y negativa (No Fraude)
  positive <- sum(performance_data$observed==1)
  negative <- sum(performance_data$observed==0)
  predicted_positive <- sum(performance_data$predicted=="Yes")
  predicted_negative <- sum(performance_data$predicted=="No")
  total <- nrow(performance_data)
  data.frame(positive, negative,predicted_positive,predicted_negative)

  # Clasificacmos a cada sujeto como éxito o fracaso
  clasificado1 <- 1*(predicted_value>=prob_fraude)
  tabla1 <- table(train$Fraude,clasificado1)
  tabla1

  #Métricas obtenidas de la matriz de confusión
  tp = sum(performance_data$observed==1 & performance_data$predicted=="Yes")
  tn = sum(performance_data$observed==0 & performance_data$predicted=="No")
  fp = sum(performance_data$observed==0 & performance_data$predicted=="Yes")
  fn = sum(performance_data$observed==1 & performance_data$predicted=="No")
  data.frame(tp,tn,fp,fn)

  accuracy =(tp+tn)/total
  error_rate = (fp+fn)/total
  sensitivity =tp/positive
  especificity = tn/negative
  precision = tp/predicted_positive
  npv = tn /predicted_negative #negative prediction value
  data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)


  PRUB <- cbind(prob_fraude,accuracy,error_rate,sensitivity,especificity,precision,npv, tp,tn,fp,fn)
  prueba[[count]] <- PRUB
  count <- count +1

}
```

```
# convertir la prueba en un único dataframe
pruebaf<-list()
pruebaf<-prueba[[1]]

for (i in 2:length(prueba)){
  pruebaf<-rbind(pruebaf,prueba[[i]])
}
pruebaf<-as.data.table(pruebaf)

head(pruebaf)
#Para definir el punto de corte de las probabilidades que optimiza los resultados se utilizan los costos
#asociados a los fallos del modelo es decir, cuanto cuesta un falso positivo (investigar un expediente
#que no resultara en fraude) vs.#cuanto cuesta un falso negativo (no investigar el expediente y el fraude
#se materializa)

#para esto se utiliza un caso base para estudio académico:
  #costes medios de fraude en el ramo de Hogar: 200
  #costos medios por cada expediente investigado: 1000

C_Inv=-200
C_Fr=-1000

pruebaf<-cbind(pruebaf, perdida=(C_Fr*pruebaf$fn)+(C_Inv*pruebaf$fp))

max(pruebaf$perdida)

pruebaf[perdida==max(pruebaf$perdida),][1] #tabla completa con probabilidad óptima

prob_fr<-pruebaf[perdida==max(pruebaf$perdida),][1]$prob_fraude
prob_fr # [1] 0.18 este es el valor de la probabilidad que optimiza el modelo en funcion de los costos


# MATRIZ DE CONFUSIÓN OBTENIDA - TRAIN
#------------------------------------------------------------

predicted_class <- ifelse(predicted_value>prob_fr, "Yes","No")
performance_data<-data.frame(observed=train$Fraude,
                             predicted= predicted_class)

#Construccion de la matriz de confusión

#Totales casos reales y predichos para respuesta positiva (Fraude) y negativa (No Fraude)
positive <- sum(performance_data$observed==1)
negative <- sum(performance_data$observed==0)
predicted_positive <- sum(performance_data$predicted=="Yes")
predicted_negative <- sum(performance_data$predicted=="No")
total <- nrow(performance_data)
data.frame(positive, negative,predicted_positive,predicted_negative)

#Métricas obtenidas de la matriz de confusión
tp = sum(performance_data$observed==1 & performance_data$predicted=="Yes")
tn = sum(performance_data$observed==0 & performance_data$predicted=="No")
fp = sum(performance_data$observed==0 & performance_data$predicted=="Yes")
fn = sum(performance_data$observed==1 & performance_data$predicted=="No")
data.frame(tp,tn,fp,fn)

accuracy =(tp+tn)/total
error_rate = (fp+fn)/total
sensitivity =tp/positive
especificity = tn/negative
precision = tp/predicted_positive
npv = tn /predicted_negative #negative prediction value
data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)
```

```
#-------------------------------------------------------------
# PREDICCIÓN EN CONJUNTO DE VALIDACIÓN
#-------------------------------------------------------------

predicted_value  <-0
predicted_value <- predict(MFinal_i1,type = "response",newdata= test)

#Se convierte el valor de predicciones en una tabla
predictor=as.table(predicted_value)

#Se exporta a excel el vector de predicciones
write.csv(predictor, file="Vector_prediccion_TEST.csv",row.names=FALSE)

#Análisis de los resultados predichos
min <- min(predicted_value)
min
max <-max(predicted_value)
max
median(predicted_value)
mean(predicted_value)

names(test)
head( predicted_value)
table(test$Fraude)
str(test$Fraude)


# CURVA ROC - TEST
#-----------------------

# require(AUC)
#curva ROC (1)
test$prob_fraude <- predicted_value
g <- roc(test$Fraude ~ test$prob_fraude , data = test)
plot(g, main="ROC - Conjunto validación")
abline(v=(0.5))


#AUC
AUC <- auc(test$Fraude, predicted_value)
AUC[1] #  0.6398284


#-----------------------------------------------------------------------------
# MATRIZ DE CONFUSIÓN EN EL CONJUNTO DE VALIDACIÓN
#-----------------------------------------------------------------------------

predicted_class<-0
predicted_class <- ifelse(predicted_value>prob_fr, "Yes","No")
performance_data<-data.frame(observed=test$Fraude,
                             predicted= predicted_class)


#Construcción de la matriz de confusión - TRAIN
#-----------------------------------------------

#Totales casos reales y predichos para respuesta positiva (Fraude) y negativa (No Fraude)
positive <- sum(performance_data$observed==1)
negative <- sum(performance_data$observed==0)
predicted_positive <- sum(performance_data$predicted=="Yes")
predicted_negative <- sum(performance_data$predicted=="No")
total <- nrow(performance_data)
data.frame(positive, negative,predicted_positive,predicted_negative)

#Métricas obtenidas de la matriz de confusión
tp = sum(performance_data$observed==1 & performance_data$predicted=="Yes")
tn = sum(performance_data$observed==0 & performance_data$predicted=="No")
fp = sum(performance_data$observed==0 & performance_data$predicted=="Yes")
fn = sum(performance_data$observed==1 & performance_data$predicted=="No")
data.frame(tp,tn,fp,fn)
```

```
accuracy =(tp+tn)/total
error_rate = (fp+fn)/total
sensitivity =tp/positive
especificity = tn/negative
precision = tp/predicted_positive
npv = tn /predicted_negative #negative prediction value
data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)
```