Trabajo Fin de Master en Estadística Aplicada

Universidad de Granada

Un modelo de regresión logística para el análisis de los aspectos que influyen en la anulación de pólizas de seguros de automóviles

María Ortuño Roig

Tutor(es): Manuel Escabias Machuca y Ana María Aguilera del Pino

Departamento de Estadística e Investigación Operativa

28 de Junio de 2022

Índice

1 Resumen

La necesidad de comprender los factores que subyacen en las causas de renovación/anulación de una póliza de seguros de automóviles está aumentando debido a la alta competencia que hay en el mercado.

El objetivo de este trabajo es hacer una revisión de la regresión logística binaria; una de las técnicas de estadística multivariante que se aplica en problemas de clasificación. Esta técnica permite estimar la relación existente entre la variable dependiente, en particular dicotómica (éxito/fracaso), y un conjunto de variables independientes.

Además, se plantea una aplicación real del uso de la regresión logística para realizar una clasificación supervisada de anulación en el mundo asegurador, utilizando un conjunto de datos de una correduría especializada en motos.

2 Introducción

Los modelos de regresión logística forman parte de los modelos lineales generalizados introducidos por Nelder y Wedderburn (1972) [19].

La regresión logística es un algoritmo de clasificación que se basa en el concepto de probabilidad. Se utiliza para modelar una variable de respuesta cualitativa en función de variables predictoras cualitativas y cuantitativas. Si la variable respuesta tiene dos clases, recibe el nombre de regresión logística binaria, y si hay más de dos clases, se denomina regresión logística múltiple. Este trabajo se centrará en la regresión logística binaria.

Se hará una revisión de dicho modelo de clasificación, desde su formulación hasta la evaluación del mismo. Además, se explicarán técnicas para elegir entre varios modelos ajustados, para así decantarnos por el que mejores resultados aporte. También se comentará brevemente el problema de los conjuntos de datos con variable dependiente no balanceada y cómo solucionarlo.

Estos modelos tienen aplicaciones en diversas áreas como la medicina, sociología, aseguradoras, marketing, banca, etc. En este trabajo se utilizará un conjunto de datos del sector asegurador para poner en práctica todos los aspectos teóricos que se explicarán sobre la regresión logística binaria. El software que se usará para esta parte práctica será R.

El documento se organiza en 9 secciones:

- Sección 1: resumen.
- Sección 2: breve introducción al trabajo.
- Sección 3: se exponen otros trabajos centrados en la predicción de la anulación, ya sea en el ámbito asegurador, banca, etc, haciendo uso de métodos de regresión logística u otros modelos de clasificación.
- Sección 4: aquí se concentrarán todos los aspectos teóricos del trabajo:

formulación del modelo logístico, estudio de la bondad del ajuste, supuestos y validación del modelo, etc.

- Sección 5: se presenta el conjunto de datos seleccionado y se exponen algunas de las librerías usadas en R.
- Sección 6: se encontrarán los resultados reportados por el modelo elegido además de su validación e interpretación.
- Sección 7: se encontrará todo el análisis descriptivo del conjunto de datos, además de las pruebas realizadas hasta llegar al modelo final.
- Sección 8: se expondrán las conclusiones finales y se indicarán algunas líneas de trabajo futuro.
- Sección 9: aquí se podrá encontrar todo el código generado en R para la realización de la parte práctica del trabajo.

3 Otros trabajos

A continuación se van a presentar algunos trabajos cuya linea de estudio es similar a la que se plantea en este. Algunos tendrán como objetivo resolver un problema parecido al que se plantea y otros servirán para lineas futuras.

An analysis of customer retention and insurance claim patterns using data mining: a case study. Smith et al. (2000) [23]

En este artículo se pretende estudiar qué patrones ayudan a predecir la renovación/anulación de una póliza además de intentar comprender qué tipo de asegurados presentan más riesgos de tener un siniestro. Con esto se busca encontrar un equilibrio entre la rentabilidad y la mejora de la retención. Además, el hecho de identificar qué pólizas tienen mayor probabilidad de anular, da la posibilidad de hacer estrategias de retención como el marketing directo.

Para realizar el estudio se tomaron 20914 pólizas de vehículos cuyos tomadores fueron notificados por carta del precio de renovación. El 7.1 % de la muestra no renovó. Se acordó que hay tres factores principales que afectarían a la decisión de anular: precio, servicio y valor asegurado del automóvil. Pero tras un estudio posterior se concluyó que el servicio no afectaba a dicha decisión.

Se utilizaron 3 técnicas para la clasificación: regresión logística, árbol de decisión y redes neuronales. Fue este último el que reportó mejores resultados.

En una segunda fase se utilizó la minería de datos no dirigida para analizar los patrones de los siniestros y así identificar aquellos grupos cuyo costo es bastante alto. Finalmente se combinaron ambos resultados para así poder determinar la prima óptima para cada póliza, de forma que se equilibre la oportunidad de obtener ganancias con la necesidad de retener al cliente.

Random Forests for Uplift Modeling: An Insurance Customer Retention Case. Guelman et. al (2012) [12]

En este artículo se propone el método de árboles aleatorios para evaluar la efectividad de una estrategia de retención. Se usa el modelado incremental para presentar una medida que se puede aplicar a cada cliente en particular. Esta técnica puede predecir la magnitud de la reacción de un cliente a una determinada campaña de marketing estratégico.

Para realizar el estudio se tomaron datos de una de las principales aseguradoras canadienses que estaba interesada en diseñar una estrategia de retención. Dicha empresa generó 3 grupos aleatorios, el primer grupo recibió una carta con su precio de renovación, el segundo grupo recibió la misma carta y una llamada de cortesía y el tercer grupo no recibió nada y sirvió como grupo de control.

Los resultados evidenciaron un impacto ligeramente positivo en el segundo grupo. Sin embargo, con el método de árboles aleatorios, se puede desentrañar si todos los clientes reaccionan de la misma manera a las campañas. Permite predecir cuánto reacciona un individuo a la persuasión de la empresa. La principal ventaja del método que se propone aquí es que permite predecir el cambio esperado en la probabilidad de que un cliente cambie a otra empresa cuando la empresa se le acerca activamente antes de la renovación. El cambio esperado se obtiene individualmente y se puede comparar con la probabilidad real prevista de cancelación de la póliza. Este enfoque conduce a una selección eficiente de los clientes a los que la empresa debe dedicar los esfuerzos de marketing.

Generalized linear models in life insurance: Decrements and risk factor analysis under solvency II. Roberto (2008) [5]

En dicho documento se analizan los datos de anulación de una importante banca aseguradora italiana. La investigación abarca el período de 1991 a 2007 con más de 6M de registros (póliza y exposición).

En el modelo resultante influyen el año natural de exposición de la póliza, los años de vigencia de la póliza en la compañía y el producto.

Los resultados del estudio reafirman la influencia en la duración de la póliza, además de lo sensible que es el modelo al año actual de exposición de la póliza. Como lineas futuras se plantea perfeccionar el modelo con datos relativos al cliente.

Predicting customer retention and profitability by using random forests and regression forests techniques. Larivière y Van den Poel (2005) [16]

En este estudio se investiga dos grupos principales de resultados de clientes: retención de clientes y rentabilidad. Se analizan dos medidas de retención que implican una transacción "activa" del cliente: la apertura de un nuevo producto (próxima compra) y la decisión de finalizar un producto que aún está abierto (deserción parcial activa). Además, también se investiga cómo evolucionan los clientes en términos de la rentabilidad que representan para la empresa mediante una variable dependiente lineal (evolución de la rentabilidad) y binaria (caída de la rentabilidad).

Se analiza una muestra de 100000 clientes de una empresa belga de servicios financieros con un amplio conjunto de variables explicativas, que incluyen el comportamiento pasado del cliente, la heterogeneidad observada del cliente y algunas variables típicas relacionadas con los intermediarios.

Se observan mejoras significativas en términos de precisión de predicción al comparar los bosques aleatorios y de regresión con los modelos de regresión logística y lineal convencionales.

En este estudio, se encuentran evidencias de que las variables del comportamiento pasado del cliente juegan un papel importante en la predicción del comportamiento futuro del cliente y la rentabilidad. Otro hallazgo importante del estudio es la importancia relativa de las variables relacionadas con los intermediarios con respecto a la clasificación de deserción parcial activa. Es claro que los buenos agentes vendedores no solo generan más compras repetidas,

sino que indirectamente evitan que los clientes hagan una deserción parcial. Por otro lado, el género del cliente y los datos geográficos recopilados a nivel de lugar de residencia son menos poderosos en términos de predicción de retención de clientes y rentabilidad.

Churn modeling of life insurance policies via statistical and machine learning methods – Analysis of important features. Groll et al. (2022) [11]

Este manuscrito trata sobre una serie de métodos de estimación estadísticos y de aprendizaje automático que intentan investigar el comportamiento de caducidad de los contratos individuales de la cartera de una gran aseguradora de vida alemana.

Se llega a la conclusión de que la base de datos más adecuada es la que contiene, además de la información sobre la póliza de seguro principal, la información sobre todas las pólizas de seguro complementarias asociadas.

Para la predicción se usaron varios modelos como GLM, Random Forest, XGBoost,... con la intención de encontrar el mejor modelo con base en estrategias de ajuste. Se tuvo en cuenta la tasa de sobremuestreo ya que los contratos de seguros de vida rara vez se cancelan y, por lo tanto, había pocos contratos cancelados en comparación con los no cancelados. Este desequilibrio presente en el conjunto de datos se tuvo en cuenta mediante el uso de un enfoque de sobremuestreo, el llamado sobremuestreo aleatorio.

Ninguno de los métodos de clasificación investigados en el proyecto superó sustancialmente a los demás ni logró una precisión predictiva muy satisfactoria.

El concepto de relevancia variable, que se utilizó para analizar e interpretar el comportamiento individual de cancelación del contrato, finalmente mostró que algunos componentes temporales, como el inicio del contrato, la duración restante del contrato o la edad del primer asegurado, pero también ciertos contenidos del contrato, como la suma asegurada, las primas anuales pagadas, el sistema de excedentes utilizado, así como la información del sistema

de recaudación, principalmente el número de amortizaciones, influyeron sustancialmente en la predicción. Dado que los valores de la bondad de ajuste del modelo no son ni particularmente buenos ni malos, estos hallazgos deben verse con cierta cautela.

4 Marco teórico

4.1 Correlación entre dos variables

Se va a hacer un breve recordatorio de los coeficientes de correlación, ya que en la parte práctica se analizará la correlación entre las variables.

Un coeficiente de correlación mide el grado en que dos variables tienden a cambiar al mismo tiempo. El coeficiente describe tanto la fuerza como la dirección de la relación. Dependiendo del tipo de variables a comparar, usaremos un coeficiente u otro:

- Coeficiente de correlación de Pearson (1895) [20]: evalúa la relación lineal entre dos variables continuas. Una relación es lineal cuando un cambio en una variable se asocia con un cambio proporcional en la otra variable. Se define el coeficiente de correlación de Pearson como:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  Siendo:

  - n = tamaño de la muestra
  - x_i, y_i = puntos muestrales individuales
  - $\bar{x}, \bar{y}$ = media muestral

- Coeficiente de correlación de Spearman (1961) [24]: evalúa la relación monótona entre dos variables continuas u ordinales. En una relación monótona, las variables tienden a cambiar al mismo tiempo, pero no necesariamente a un ritmo constante.

  La correlación de Spearman suele utilizarse para evaluar relaciones en las que intervienen variables ordinales y es muy útil cuando el número de pares que se desea asociar es pequeño (menor de 30).

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Siendo:

- n = cantidad de sujetos que se clasifican
- x_i = rango de sujetos i con respecto a la variable x
- y_i = rango de sujetos i con respecto a la variable y
- $d_i = x_i - y_i$ es la diferencia entre los rangos X e Y.

El hecho de que las dos variables tiendan a crecer o decrecer juntas no indica que una tenga efecto directo o indirecto sobre la otra. Ambas pueden estar influidas por otras variables de modo que se origine una fuerte relación entre ambas. La interpretación de este coeficiente depende principalmente de la experiencia propia en el tema de estudio.

Ambos coeficientes pueden tomar valores entre -1 y 1. El valor 0 indica que no existe asociación lineal entre ambas variables. Puede que exista otro tipo de correlación, pero no lineal. Los signos positivos o negativos solo indican la dirección de la relación.

Es necesario calcular la significancia de las correlaciones obtenidas. Cuando el valor p de significación es menor que 0.05, se puede concluir que la relación es significativa a un nivel de significación $\alpha = 0.5$

4.2 Conjuntos de datos no balanceados

En los problemas de clasificación binaria (o múltiple) se suele trabajar con conjuntos de datos que no están balanceados respecto a la variable respuesta.

Los conjuntos de datos en los que la variable respuesta no está balanceada, corren el riesgo de que los modelos de predicción que se le apliquen acaben dando resultados sesgados, a favor de la clase mayoritaria. Esto significa que

las observaciones de la clase minoritaria a menudo se predicen incorrectamente, mientras que la tasa general de clasificación errónea sigue siendo baja.

Existen varias técnicas para corregir esto:

- Resampling: esta técnica se utiliza para aumentar o reducir la muestra de la clase minoritaria o mayoritaria.
  - Oversampling: sobremuestrea la clase minoritaria usando reemplazo.
  - Undersampling: elimina registros aleatorios de la muestra mayoritaria hasta que ambas clases tienen la misma cardinalidad.
- ROSE: usa bootstrapping suavizado para extraer muestras artificiales de la clase minoritaria basándose en los vecinos más cercanos.
- SMOTE: técnica para sobremuestrear la clase minoritaria. Es la solución para el sobreajuste. Supera el problema generando datos artificiales, en lugar de replicar o agregar las observaciones de la clase minoritaria. Se puede encontrar mas información en Chawla et al. (2002) [6].

Estos métodos tratan de cambiar la estructura de los datos durante el entrenamiento de los modelos, con el objetivo de conseguir un conjunto de datos de entrenamiento equilibrado con respecto a la variable respuesta.

4.3 El modelo logístico

4.3.1 Introducción

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite explicar el comportamiento de una variable respuesta binaria o múltiple (variable dependiente) sobre la base de uno o varios predictores de naturaleza cuantitativa y/o cualitativa.

El modelo de respuesta binaria es la forma más simple de regresión logística, en la cuál, la variable respuesta solo puede tomar dos valores, 1 (éxito) o 0

(fracaso). También se puede pensar que 1 hace referencia a tener alguna propiedad, condición o característica y 0 que no tiene esa propiedad, condición o característica.

Los modelos de respuesta discreta son un caso particular de los modelos lineales generalizados formulados por Nelder y Wedderburn (1972) [19].

Los dos principales usos del modelo de regresión logística son:

- Interpretar las estimaciones de los parámetros del modelo.
- Calcular la probabilidad de que la variable respuesta tome el valor 1 (éxito).

Ambos usos juegan un papel importante en múltiples áreas como son: salud e investigación médica, calificación crediticia, investigación en ciencias sociales, etc.

A continuación se planteará la formulación genérica del modelo de regresión logística binaria, así como la estimación e interpretación de los parámetros, terminando por la evaluación y selección del mejor modelo. Aunque para un estudio exhaustivo se recomienda la consulta de otros materiales como Hosmer y Lemeshow (2013) [14], Nelder Y Wedderburn (1972) [19] o Collett (2002) [7].

4.3.2 Formulación

Para el objetivo mencionado anteriormente de predecir el valor de la variable respuesta ya sea entre dos niveles (binario) o más niveles (múltiple) se podría pensar en usar la Regresión Lineal Simple o Múltiple (dependiendo del caso), pero esto no será posible ya que el valor que esperamos obtener es una probabilidad (de pertenecer a un nivel, o a otro, en la variable respuesta).

En la Figura 1 se puede ver como si se utilizara la regresión lineal se obtendrían valores superiores a 1 e inferiores a -1, por lo que no podría interpretarse el resultado como una probabilidad.

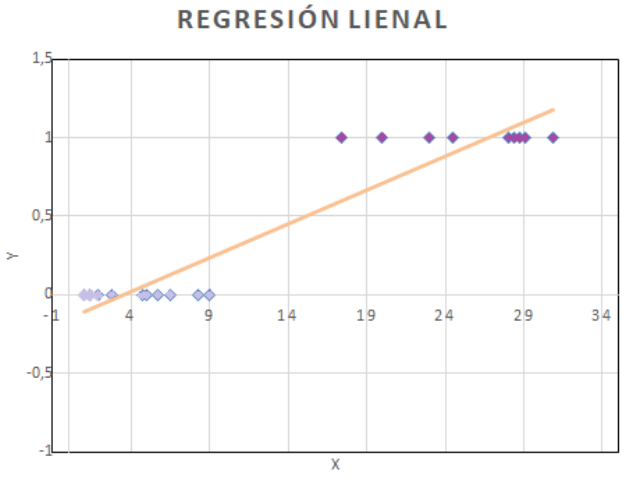


Figura 1: Regresión lineal.

Para evitar el problema anterior, la regresión logística transforma el valor devuelto por la regresión lineal ($Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$) empleando una función cuyo resultado esté siempre comprendido entre 0 y 1.

Aunque existen varias funciones que cumplen esta condición, una de las más utilizadas es la función logística o sigmoide:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

La función anterior cumple que $\lim_{x \to -\infty} \sigma(x) = 0$ y $\lim_{x \to \infty} \sigma(x) = 1$.

Sustituyendo en la función sigmoide la x por la función lineal $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$ se obtiene que

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}}$$

Los valores posibles para esta ecuación van desde 0 hasta 1, lo que hace que pueda ser interpretado como una probabilidad. Cuanto más cercano esté el valor de P(Y) a 1, más probable es que suceda Y.

Supongamos que tenemos una variable respuesta Y que toma dos valores (1 o 0) y n variables predictoras representadas por $X = (X_1, X_2, ..., X_n)'$. La formulación genérica del modelo de regresión logística para modelar la probabilidad de ocurrencia de un suceso Y es $Y = p_x + \epsilon$ donde ϵ es el término de error y p_x es la probabilidad de que la respuesta Y tome el valor 1 para el valor observado x. Se modeliza como:

$$P(Y = 1|X = x) = p_x = \frac{e^{\beta_0 + \sum_{i=1}^{n} b_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} b_i x_i}} \tag{4.1}$$

donde:

- $P(Y = 1|X = x)$ es la probabilidad de que Y tome el valor 1 para el valor observado x
- $1 - P(Y = 1|X = x)$ es la probabilidad de que Y tome el valor 0 para el valor observado x
- X es un conjunto de n variables predictoras $(x_1, ..., x_n)$ que forman parte del modelo
- b_0 es la constante del modelo o término independiente
- b_i son los coeficientes de las variables predictoras

Si se divide la expresión 4.1 por su complementario, se obtiene lo que se conoce como el odds de la variable respuesta:

$$\frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)} = e^{\beta_0 + \sum_{i=1}^{n} b_i x_i} \tag{4.2}$$

Como la ecuación 4.2 no es sencilla de interpretar, se puede aplicar el logaritmo para así obtener el llamado logit. Es decir, la función logit es el logaritmo natural del odds de la variable respuesta.

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \beta_0 + \sum_{i=1}^{n} b_i x_i \tag{4.3}$$

Así se obtendría la expresión de una recta, idéntica a la del modelo general de regresión lineal.

Para ampliar información se puede consultar el libro "Building and Applying Logistic Regression Models" de Agresti (2003) [1].

4.3.3 Supuestos del modelo

Los modelos de regresión logística deben verificar los siguientes supuestos:

- Variable respuesta binaria
- Linealidad: el supuesto de linealidad en regresión logística se cumple si existe una relación lineal entre cada variable predictora y el logaritmo de la variable respuesta. La forma más fácil de evaluar esta suposición es usando la prueba de Box-Tidwell (1962) [4].
- Observaciones independientes: las observaciones no deben venir de mediciones repetidas del mismo individuo ni estar relacionadas entre sí de ninguna manera. Para verificar esta suposición se creará un gráfico de residuos contra el tiempo (es decir, el orden de las observaciones) y se observará si existe o no un patrón aleatorio. Si no hay un patrón aleatorio, entonces este supuesto puede que no se esté cumpliendo.
- Multicolinealidad: las variables predictoras no deben estar altamente correlacionadas. La forma más común de detectar la multicolinealidad es mediante el uso del factor de inflación de la varianza (VIF).

  El VIF asociado a cada variable predictora del modelo mide la relación entre la varianza general del modelo y la varianza del modelo que incluye sólo esa variable predictora. Un VIF alto asociado a una variable predictora indica que dicha variable es altamente colineal con el resto de variables predictoras del modelo.

Detectar la multicolinealidad es importante porque aunque no reduce el poder explicativo del modelo, sí reduce la significación estadística de las variables predictoras.

Para saber cómo se calcula y cómo se interpreta, se propone consultar Kleinbaum (2013) [15].

- No hay valores atípicos extremos: para verificar esta suposición se calculará la distancia de Cook para cada observación.

  La distancia de Cook se calcula eliminando el i-ésimo registro del modelo y volviendo a ajustar el modelo. Es decir, resume cuánto cambian todos los valores del modelo de regresión cuando se elimina el i-ésimo registro.

  En el caso de existir valores atípicos se optará por eliminarlos, reemplazarlos por un valor como la media o mediana o simplemente mantenerlos en el modelo pero se tomará nota de esto al informar los resultados de la regresión.

  Para ampliar información se puede consultar Boussiala (2020) [3].

- Muestra grande: se asume que el tamaño de la muestra del conjunto de datos es lo suficientemente grande como para sacar conclusiones válidas del modelo de regresión logística ajustado.

4.3.4 Contrastes sobre los parámetros del modelo

El estadístico de Wald (z-statistic) proporciona la contribución individual de cada uno de los parámetros del modelo.

Para contrastar si un parámetro es significativo o no, se tendrá que realizar el test de significación de parámetros (coincidirá con el test de significación de variables en el caso de variables cuantitativas):

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases} \tag{4.4}$$

El contraste a realizar, evalúa si el parámetro asociado a dicha variable es igual a cero o no.

Para el parámetro i-ésimo, el estadístico de Wald es

$$z = \frac{\hat{\beta}_i}{S\hat{E}_{\beta_i}}$$

siendo $\hat{\beta}_i$ las estimaciones máximo verosímiles de βi y $S\hat{E}_{\beta_j}$ las estimaciones de su correspondiente desviación estándar.

Se verifica que $z \to N(0,1)$ o lo que es equivalente, $z^2 \to \chi^2_1$.

Por lo tanto, se rechazará la hipótesis nula al nivel de significación α cuando se verifique que

$$z^2 \leq \chi_{1;\alpha}$$

Si es distinto de cero asumimos que dicha variable predictora está haciendo una contribución significativa al modelo para predecir el valor de Y y por tanto se debe mantener.

4.3.5 Interpretación de los coeficientes

Recordemos que la ventaja de respuesta Y = 1 para el valor observado X = x, se conoce como **odds** y viene dado por el cociente

$$odds_x = \frac{p_x}{1 - p_x}$$

Esto significa que el odds de un suceso Y es el cociente de la probabilidad de ocurrencia de dicho suceso entre la probabilidad de no ocurrencia, bajo unas determinadas condiciones x.

Por otro lado, el **cociente de ventajas (odds ratio)** de respuesta Y = 1 dados

dos valores distintos x_1 y x_2 de la variable explicativa X, es de la forma

$$\theta_{1,2} = \frac{\frac{p_{x_1}}{1-p_{x_1}}}{\frac{p_{x_2}}{1-p_{x_2}}}$$

Podemos interpretar el cociente de ventajas de la siguiente forma:

- Si es mayor que 1 indica que si el predictor aumenta, el odds de la variable dependiente crece.
- Si es menor que 1 indica que si el predictor aumenta, el odds de la variable dependiente decrece.

Existe una relación entre probabilidades y odds: una probabilidad puede convertirse en odds mediante la fórmula $\frac{probabilidad}{1-probabilidad}$, y un odds convertirse en una probabilidad mediante la fórmula $\frac{odds}{odds+1}$.

Riesgo	Odds
0,1	0,1/0,9 = 0,11
0,2	0,2/0,8 = 0,25
0,3	0,3/0,7 = 0,43
0,4	0,4/0,6 = 0,67
0,5	0,5/0,5 = 1,00
0,6	0,6/0,4 = 1,50
0,7	0,7/0,3 = 2,33
0,8	0,8/0,2 = 4,00
0,9	0,9/0,1 = 9,00

Riesgo = (*odds*/(*odds*+1)). Odds = (riesgo/(1-riesgo)). Los riesgos toman valores entre 0 y 1, y los *odds* entre 0 e infinito. A mayor magnitud de riesgo, mayor es la diferencia numérica con su respectivo *odds*. Fuente: revista médica Chile

Figura 2: Riesgo vs. Odds

A partir de la formulación del modelo se tiene de manera inmediata que:

- β_0: corresponde con el logaritmo del odds de Y = 1 frente a la respuesta Y = 0 cuando la respuesta es independiente de las variables explicativas.

$$P(Y = 1|X = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$\frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)} = e^{\beta_0}$$

$$\log(\frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)}) = \beta_0$$

- β_i: su exponencial es el cociente de ventaja de respuesta Y = 1 para dos observaciones de la variable explicativa que se diferencian en una unidad.

$$\theta(\triangle X_1 = 1, ..., \triangle X_n = 1) = e^{\sum_{r=1}^{n} \beta_r} = \prod_{r=1}^{n} e^{\beta_r}$$

Si la diferencia es en más de una unidad:

$$\theta(x_1, x_2) = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} = e^{\sum_{r=1}^{n} \beta_r (x_{1r} - x_{2r})}$$

donde $x_1 = (1, x_{11}, ..., x_{2n}) y x_2 = (1, x_{21}, ..., x_{2n})$.

4.3.6 Bondad de ajuste

En esta sección se van a mostrar distintas formas de medir cómo de bueno es el ajuste de los valores predichos por el modelo, a los valores observados.

Las medidas de bondad de ajuste, en general, resumen la discrepancia entre los valores observados y los valores esperados en el modelo de estudio.

De forma global, puede ser evaluada a través de medidas tipo R^2, la tasa de clasificaciones correctas,... Se estudiarán varios de estos test estadísticos de bondad de ajuste.

El contraste de hipótesis a realizar será:

$$\begin{cases} H_0: & p_q = \frac{e^{\sum_{i=0}^{n} \beta_i x_{qi}}}{1+e^{\sum_{i=0}^{n} \beta_i x_{qi}}}, \ \forall q = 1, ..., m \\ H_1: & p_q \neq \frac{e^{\sum_{i=0}^{n} \beta_i x_{qi}}}{1+e^{\sum_{i=0}^{n} \beta_i x_{qi}}}, \ para\ algun\ q \end{cases} \tag{4.5}$$

Devianza o estadístico de Wilks

Como se ha comentado anteriormente, el modelo logístico devuelve para cada individuo, la probabilidad de que ocurra el suceso Y, es decir, la probabilidad de que Y tome el valor 1.

Por tanto, para cada individuo se tendrá tanto su valor observado para la variable Y como el valor predicho por el modelo. Se usará ambos valores para evaluar el ajuste del modelo.

Sea $\hat{L}_{ajustado}$ la función de verosimilitud del modelo ajustado, se define el log-likelihook (logaritmo de la función de verosimilitud) como:

$$log - likelihood = \sum_{q=1}^{m} [y_q \log p_q + (1 - y_q) \log(1 - p_q)]$$

Este estadístico se usa como indicador de cuánta información sin explicar queda en la variable respuesta tras haber ajustado el modelo. Los valores de log-verosimilitud no se pueden utilizar por sí solos como un índice de ajuste, porque dependen del tamaño de la muestra, pero sí se pueden utilizar para comparar el ajuste de diferentes modelos ya que lo que se desea es maximizar su valor.

Considérese ahora, $\hat{L}_{saturado}$, la función de verosimilitud del modelo saturado (modelo que se ajusta perfectamente a los datos). La devianza (Collett (2002) [7]) trata de comparar ambas funciones para medir la bondad del ajuste del modelo.

Se define la devianza o estadístico de Wilks como:

$$Devianza = -2\log\left(\frac{\hat{L}_{ajustado}}{\hat{L}_{saturado}}\right)$$

Dicho estadístico mide si el modelo se ajusta bien a los datos. Si la devianza es grande, el modelo no se ajusta bien a los datos. En caso contrario, sería un buen ajuste.

Pero... ¿cómo saber cual es el umbral que indica si la devianza es grande o pequeña?

La devianza sigue un a distribución asintótica Chi-Cuadrado con m-n grados de libertad, donde m es el número de datos a ajustar y n el número de parámetros del modelo. Por tanto, si la $devianza \geq \chi^2_{m-n;\alpha}$ se rechazará la hipótesis nula a un nivel de significación α y se asumirá que el modelo se ajusta bien a los datos.

Chi-Cuadrado de Pearson

El estadístico Chi-Cuadrado de Pearson compara frecuencias observadas y esperadas en una distribución binomial.

Dicho estadístico se define como:

$$\chi^2 = \sum_{q=1}^{m} r_q^2$$

donde $r_q = \frac{y_q - n_q\hat{p_q}}{\sqrt{n_q\hat{p_q}(1-\hat{p_q})}}$ denominados por Hosmer como residuos de Pearson.

El contraste a realizar para verificar su significación estadística es:

$$\begin{cases} H_0: & r_q = 0 \\ H_1: & r_q \neq 0 \end{cases} \tag{4.6}$$

Bajo la hipótesis nula r_q tiene distribución asintótica normal con media cero y varianza estimada menor que uno. A pesar de esto los residuos de Pearson suelen ser tratados como normales estándar. Se considerarán significativos cuando sus valores absolutos sean mayores que dos.

Para ampliar información se puede consultar Plackett (1983) [21].

R y R^2

En regresión logística podemos calcular una medida análoga al R^2 que se usa en el modelo de regresión lineal. Esta medida se conoce como "pseudo R cuadrado".

Son varias las medidas que se han propuesto para asemejarse al R^2, pero en esta sección se explicará solo la de MacFadden (1973) [17]:

$$R_L^2 = |\frac{2LL(completo) - 2LL(nulo)}{2LL(nulo)}|$$

Es la reducción proporcional en valor absoluto de log-likelihood y mide cuánto del error del ajuste disminuye al incluir las variables predictoras. Proporciona una medición de la significación real del modelo.

- Se trata de la correlación parcial entre la variable respuesta y cada una de las predictoras.
- Puede variar entre 0 y 1.
- Un valor cercano a 1 significa que al crecer la variable predictora, lo hace la probabilidad de que el evento ocurra.
- Un valor cercano a 0 implica que si la variable predictora decrece, la probabilidad de que el resultado ocurra disminuye.
- Si una variable tiene un valor pequeño de R_L^2 contribuye al modelo sólo en una pequeña cantidad.

Otras medidas que podrían ser útiles son las de Tjur (2009) [25], Cox (2018) [9] y Nagelkerke (1991) [18].

Tasa de clasificaciones correctas

Es la proporción de registros clasificados correctamente por el modelo. Un registro es clasificado correctamente por el modelo logístico cuando el valor real de la variable respuesta coincide con su valor estimado por el modelo.

Para poder categorizar una observación en el valor Y = 1 o Y = 0 tras aplicar el modelo, se elige un punto de corte, $p \in (0,1)$, de modo que a una observación se le asigna respuesta Y = 1 si $p_q > p$ y se le asigna respuesta Y = 0 cuando $p_q \leq p$.

¿Qué criterios usar para elegir dicho punto de corte? Se suele elegir 0.5 aunque en otras ocasiones, es más apropiado elegir la proporción de unos en la muestra o elegirlo en base al problema que intentamos resolver, ya que en algunas ocasiones se necesitará ser más restrictivo con la clasificación positiva (1) y otras no tanto.

Curva ROC

Dada una tabla de contingencia de la forma:

	Valor en la realidad	
Predicción	Verdaderos Positivos	Falsos Positivos
	Falsos Negativos	Verdaderos Negativos

Figura 3: Tabla contingencia.

La capacidad predictiva de un modelo de regresión logística se puede resumir mediante el concepto de sensibilidad:

$$sensibilidad = \frac{VP}{VP + FN}$$

y mediante el concepto de especificidad:

$$especificidad = \frac{VN}{VN + FP}$$

Es decir, la probabilidad de predecir correctamente el ”éxito” de la variable respuesta, se denomina sensibilidad y la probabilidad de predecir correctamente el ”fracaso” de la variable respuesta se denomina especificidad.

El cálculo de estos valores es muy sensible a cuales son las frecuencias relativas de y = 1 e y = 0.

Una curva de tipo receiver operating characteristic (ROC) es un gráfico en el que se representa la sensibilidad en función de (1 – especificidad). Si se modifica los valores del valor de corte C seleccionado para clasificar cada registro en Y = 1 o Y = 0 y representamos la sensibilidad (en ordenadas) frente a (1 – especificidad) (en abscisas) tenemos la curva ROC. Es una curva cóncava que conecta los puntos (0, 0) y (1, 1). Cuanto mayor sea el área bajo la curva mejores serán las predicciones.

La curva ROC ofrece un mejor resumen de la capacidad predictiva que una tabla de clasificación, porque presenta la potencia predictiva para todos los posibles valores de referencia C.

El área que se encuentra bajo la curva se denomina AUC. Se pueden dibujar varias curvas ROC de distintos modelos para comparar su potencia de predicción.

4.3.7 Evaluación

En la práctica, antes de ajustar el modelo, se separará el conjunto de datos en dos subconjuntos, uno de entrenamiento y otro de prueba. Con el primero de ellos se entrenará el modelo y con el segundo se validará el mismo.

El conjunto de prueba debe cumplir que:

- Es lo suficientemente grande como para producir resultados estadísticamente significativos
- Es representativo del conjunto de datos global.

El objetivo es crear un modelo que generalice bien los nuevos datos y que no esté sobreajustado, es decir que solo se ajuste bien al conjunto de datos de prueba.

Teniendo en cuenta lo anterior, se puede entrar a explicar cómo evaluar un modelo.

Para evaluar cómo se ajusta el modelo a los datos de prueba, se comparará la clasificación predicha por el modelo con su valor real. De esta forma se calculará la tasa de clasificaciones correctas (ver sección 4.3.6).

Además, se examinarán los residuos para asegurar que el modelo se ajusta bien a los datos observados (ver Cook (1982) [8]). Para esto último se buscará:

- Aislar los puntos en los que el modelo se ajusta mal.
- Aislar los puntos que ejercen una influencia excesiva sobre el modelo.

Se podrá detectar los casos conflictivos de la siguiente forma:

- No más del 5 % de los residuos estandarizados tiene un valor absoluto mayor de 2, y que no más de un 1 % tiene valores absolutos más allá de 2.5. Cualquier caso con valor superior a 3 podría ser un valor atípico.

Se utilizarán los residuos estandarizados para poder compararlos entre sí, ya que estarán en una misma escala. Se definen los residuos estandariados como:

$$r_q^s = \frac{r_q}{\sqrt{1 - h_{qq}}}$$

donde h_{qq} es el elemento diagonal de la matriz

$$H = W^{\frac{1}{2}} X (X'WX)^{-1} X' W^{\frac{1}{2}}$$

con W = $Diag[n_q \hat{p}_q (1 - \hat{p}_q)]$

- El valor h_{qq} del punto anterior describe la influencia de la observación x_q. Si h_{qq} es pequeña, entonces la respuesta observada tiene poca influencia en $\hat{y}_q$. Por el contrario, si es grande, tiene una gran influencia en $\hat{y}_q$. Como regla general, las observaciones con un alto apalancamiento (leverage) son aquellas cuyo valor es superior a $\frac{2n}{q}$ donde n es el número de predictores y q el número de observaciones.

Identificar estos valores conflictivos no es suficiente razón para eliminarlos del modelo. Una vez identificados, se debe analizar estos casos para encontrar el motivo por el que son inusuales. Una vez hecha la investigación, si por ejemplo encontramos que es un valor erróneo, sí podríamos proceder a eliminarlo del modelo.

4.3.8 Comparación y selección del modelo

AIC y BIC

Para comparar el ajuste entre dos modelos se usará el criterio de información de Akaike (AIC) y el criterio de información de Bayes (BIC). Ambos criterios fueron propuestos por Akaike (1974) [2] y Scharwarz (1978) [22] respectivamente. Estos criterios proporcionan una medida del ajuste de un modelo que penaliza al modelo que contiene más variables predictoras.

$$AIC = -2LL + 2k$$

$$BIC = -2LL + 2k\log(n)$$

Al ajustar un modelo es posible aumentar su precisión mediante la adición de parámetros, pero si esto se lleva acabo, se puede derivar en un sobreajuste. Tanto el BIC y AIC resuelven este problema mediante la introducción de un término de penalización para el número de parámetros en el modelo. Este término de penalización es mayor en el BIC que en el AIC.

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC y BIC.

Métodos paso a paso

El objetivo subyacente de la regresión paso a paso es, a través de una serie de test de hipótesis (p. ej., pruebas F, pruebas t), encontrar un conjunto de variables independientes que influyan significativamente en la variable dependiente. Esto se hace con programas informáticos a través de la iteración. La realización de pruebas de forma automática con la ayuda de paquetes de software estadístico tiene la ventaja de ahorrar tiempo y limitar los errores.

La regresión paso a paso se puede hacer partiendo del modelo nulo e ir probando cada variable independiente de forma que esta se incluye en el modelo si es estadísticamente significativa o, en cambio, partiendo del modelo con todas las variables e ir eliminando aquellas que no son estadísticamente significativas.

También se podría usar una combinación de ambos métodos y, por lo tanto, habría un total de tres formas de realizar la regresión paso a paso:

- La selección hacia adelante comienza sin variables en el modelo, prueba cada variable a medida que se agrega al modelo y luego conserva las que se consideran estadísticamente más significativas (buscando siempre que

se mejore el AIC o BIC), repitiendo el proceso hasta que los resultados sean óptimos.

- La eliminación hacia atrás comienza con un conjunto de variables independientes, eliminando una a la vez y luego probando para ver si la variable eliminada es estadísticamente significativa (buscando siempre que se mejore el AIC o BIC).
- La eliminación bidireccional es una combinación de los dos primeros métodos que prueban qué variables deben incluirse o excluirse (buscando siempre que se mejore el AIC o BIC).

Se sugiere leer Efroymson (1969) [10] y Hocking (1976) [13] para ampliar información.

Otros

Algunas de las medidas de bondad de ajustes comentadas en apartado 4.3.6 se pueden usar también para comparar modelos en los que ninguno de ellos es el modelo nulo. Por ejemplo se puede usar la devianza o el criterio de máxima verosimilitud. También se podría comparar las curvas ROC de cada modelo.

4.3.9 Resumen: etapas de la regresión logística

- Análisis descriptivo de las variables.
- Dividir el conjunto de datos de forma aleatoria, en dos subconjuntos. Un subconjunto de entrenamiento y otro de prueba. Se deberá asegurar que la ponderación de la variable respuesta en el conjunto original se mantenga en el conjunto de entrenamiento.
- Comparar varios modelos para ver cual devuelve mejores resultados
- Comprobar que se cumplen los supuestos del modelo logístico.
- Evaluar la bondad del modelo aplicando lo comentado en la sección 4.3.6.

- Para cada parámetro del modelo, estudiar su significación haciendo uso del estadístico de Wald.
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba.
- Usar los cocientes de ventajas para interpretar el modelo. Si el valor es mayor que 1 al aumentar la variable predictora el odds de la respuesta aumenta. Inversamente, un valor menor que 1 indica que si la variable predictora crece, el odds de la respuesta decrece.
- Estudiar la potencia de predicción del modelo y ver que se ajusta bien a los datos.

5 Material y métodos

Con el propósito de poner en práctica estos conceptos, se han tomado los datos de una correduría de seguros especializada fundamentalmente en la tramitación de seguros de motos. Haciendo uso de estos datos, se tratará de predecir qué harán los clientes al vencimiento de su póliza, si anular o renovar (renovación anual).

Para poder hacer un correcto análisis de los datos es importante entender la estructura de los mismos, por lo que se quiere dejar claro antes de empezar esta sección de resultados, que los datos escogidos están presentados de forma que se dispone de un registro por póliza y años en vigor, es decir, que una póliza que haya estado 2 años en la aseguradora, aparecerá dos veces, una por el primer año y otra por el segundo. Por lo que cada una de las variables elegidas harán referencia a lo que haya sucedido en ese año en cuestión.

Con el fin de dar respuesta a la pregunta de qué hará el cliente en el momento de la renovación, se ha tenido que realizar un proceso de recopilación, limpieza y transformación de los datos. Todo esto se ha hecho con SQL Server.

No quería pasar por alto esta parte, ya que para la realización del trabajo no se ha seleccionado un conjunto de datos preprocesado y se ha tenido que emplear tiempo en la selección y limpieza de los mismos.

Una vez seleccionados los datos a tratar, la posterior carga en R se ha hecho a través de la librería ODBC de la que hablaremos mas adelante.

Antes de entrar en más detalles se realizarán unas serie de puntualizaciones sobre el conjunto de datos elegido:

- Se han elegido pólizas exclusivamente de motos, que es el grueso del negocio.
- Se han seleccionado los registros que tuvieron vigencia desde 2018 hasta marzo de 2021.

- Se han considerado perfiles de clientes estándar, es decir, se han descartado flotas de alquiler, motos demo, etc.
- Se han descartado tipos de vehículos no comunes como pueden ser motos de campo, microcars, ... donde apenas tenemos datos.
- Los registros que anularon a mitad de anualidad, han sido descartados del estudio ya que dichas anulaciones se deben a causas como robo, venta de vehículo,... cosas que no tienen que ver con la empresa directamente.

Esto lleva a una selección de 227320 registros y 44 variables, de las cuales tras realizar un estudio de cada una, finalmente se ha hecho uso únicamente de 20:

- Variables predictoras:

  1. Mes: mes de renovación.
  2. Mediador: indica a través de quién se hace el seguro.
  3. Marca: marca del vehículo.
  4. División: agrupaciones de agentes (entendiéndose agente por quién hace la venta): Kawasaki, Peugeot, SYM, ...
  5. Año de renovación: indica el número de años que el cliente ha estado en la empresa, de forma que alguien que vaya a pasar por su primera renovación, tomará el valor 1.
  6. Club: es un paquete con garantías extras. Una persona puede no tenerlo contratado, contratar su versión básica o contratar su versión plus. El paquete que contrate tiene su efecto en el precio del seguro.
  7. Tipo de vehículo: tipo de vehículo (ciclomotor, deportiva, scooter, ...)
  8. Cilindrada: cilindrada del vehículo (menos de 50 cc, 125 cc, de 126 a 250 cc,...)
  9. Accidentes: indica si tiene contratado accidentes o no. En caso afirmativo, si tiene la versión básica o la plus.

10. Asistencia en viaje: indica si tiene contratada la asistencia en viajes o no.
11. Defensa: indica si tiene contratado la defensa o no.
12. Paquete de garantías: indica el tipo de seguro contratado (terceros, robo e incendios o todo riesgo).
13. Perfil: hay definidos 4 perfiles en función de la edad y los años de carnet: 2 Recargos, 1 Recargo, Base y Bonificado. Para ello se calcula el perfil de las tres figuras, tomador, propietario y asegurado. La póliza llevará asignado el perfil más bajo de las tres figuras.
14. Nivel crediticio: nivel crediticio del tomador del seguro.
15. Edad (factor): agrupación para la edad del tomador.
16. Carnet (factor): agrupación para los años de carnet.
17. Matricula (factor): agrupación para la antigüedad del vehículo.
18. Prima total: prima total de renovación.
19. Diferencia de prima: la diferencia entre el valor de la prima total con respecto al precio del año anterior.

- Variable respuesta:

20 Estado de la póliza: vigor (0) /anulada (1) en función de la decisión del cliente al final de la anualidad.

Otras variables que también han sido analizadas pero que se descartaron finalmente son:

- Regalos: son 6 variables que hacen referencia a tipos de regalos que pueden recibir los clientes (gafas, chubasquero, cheque regalo,...). Indica si el cliente ha recibido dicho regalo (1) o no (0).
- Garantías: son 17 variables que hacen referencia a coberturas que se pueden contratar (cambio de aceite, reparación de neumáticos, vehículo de sustitución, seguro de desempleo,...). Indica si el cliente tiene contratada la cobertura o no, y si la tiene, si la ha usado o no. Tomando el valor 9 si no la tiene contratada, 0 si no la ha usado y 1 si sí la ha usado.

- Provincia: provincia del tomador del seguro.
- Grupo cartera: es una variable categórica con dos niveles (grupo de control y reglas). A las pólizas que pertenecen al grupo de control no se le aplican reglas de negocio sobre la prima.

El software elegido para la realización del trabajo ha sido el software libre R[1]. Para la realización del análisis descriptivo, selección de variables, formulación del modelo y evaluación del mismo se ha hecho uso de los siguientes paquetes:

- dplyr: se utiliza para trabajar con marcos de datos como objetos.
- caret: es un conjunto de funciones que intentan agilizar el proceso de creación de modelos predictivos. El paquete contiene herramientas para el preprocesamiento de los datos, selección de variables, ajuste de modelos, estimación de la importancia de las variables, etc.
- ggplot2: paquete de visualización de datos.
- pROC: librería para calcular el área bajo la curva ROC.
- ODBC: librería para hacer consultas a bases de datos.
- GGally: gráficos de correlaciones.
- gridExtra: para unir varios gráficos en una misma cuadrícula.
- lubridate: manipular fechas.
- car: calcular VIF y distancia de Cook.

El uso de dplyr ha sido esencial durante todo el trabajo para el manejo del conjunto de datos.

La lectura de los datos se ha hecho a través del paquete ODBC. Esto permite hacer las consultas de manera automática, con valores actualizados en caso de

[1] https://cran.r-project.org/

haber algún cambio en la base de datos original. Además, esta misma consulta permitirá la extracción de datos futuros a los que aplicar el modelo y obtener las predicciones de anulación/renovación que se están buscando.

En lo que respecta al análisis descriptivo de los datos se ha hecho uso de paquetes como ggplot para su visualización, ggally para el estudio de las correlaciones y car para estudiar los supuestos del modelo.

A la hora de particionar los datos en los conjuntos de entrenamiento y prueba se ha hecho uso del paquete caret. Y para la validación del modelo se ha usado por ejemplo, el paquete pROC.

El código empleado para generar la parte práctica se podrá encontrar en la sección 9 (Anexo) de este trabajo.

6 Resultados

En esta sección se mostrará el ajuste del modelo seleccionado, se estudiará su bondad, se revisará que se cumplen los supuestos del modelo y se evaluará el resultado obtenido.

En la sección 7 (otros aspectos prácticos) se podrá encontrar el análisis previo que se ha realizado para poder determinar cuál es el modelo que mejor se ajusta a los datos.

De todas las variables que se han expuesto anteriormente, únicamente 6 de ellas se utilizarán en el modelo final. Estas son: el **nivel crediticio**, la **diferencia de prima**, el **perfil**, **año de renovación**, **paquete de garantías** y **club** para ajustar el modelo.

```
Call:
glm(formula = ANULA ~ CreditScoring + PrimaVigorTotalDiferencia +
    Perfil + AnyoRenovacionComparativa + PaqueteGarantias + Club,
    family = "binomial", data = trainup)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7352  -0.9765  -0.0607   1.0399   3.4981

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.960e-01  4.188e-02 -23.780  < 2e-16 ***
CreditScoring5                   1.131e-01  1.869e-02   6.054 1.41e-09 ***
CreditScoring4                   2.842e-01  1.758e-02  16.168  < 2e-16 ***
CreditScoring3                   4.247e-01  1.956e-02  21.714  < 2e-16 ***
CreditScoring1 - 2               9.573e-01  3.474e-02  27.561  < 2e-16 ***
CreditScoringA                   9.342e-01  4.524e-02  20.648  < 2e-16 ***
CreditScoringB                   9.637e-01  4.836e-02  19.928  < 2e-16 ***
CreditScoringC                   9.707e-01  5.445e-02  17.826  < 2e-16 ***
CreditScoringD                   1.371e+00  7.249e-02  18.909  < 2e-16 ***
CreditScoringF - E               1.387e+00  8.545e-02  16.230  < 2e-16 ***
CreditScoring-                   1.716e+00  2.188e-02  78.398  < 2e-16 ***
PrimaVigorTotalDiferencia        6.068e-03  6.927e-05  87.591  < 2e-16 ***
Perfil1 RECARGO                  7.570e-01  4.047e-02  18.705  < 2e-16 ***
PerfilBASE                       8.246e-01  3.916e-02  21.057  < 2e-16 ***
PerfilBONIFICADO                 7.409e-01  3.739e-02  19.818  < 2e-16 ***
AnyoRenovacionComparativa       -1.094e-01  2.075e-03 -52.747  < 2e-16 ***
PaqueteGarantiasROBO E INCENDIO -3.209e-01  1.209e-02 -26.552  < 2e-16 ***
PaqueteGarantiasTODO RIESGO     -3.563e-01  1.388e-02 -25.669  < 2e-16 ***
ClubCLUB BASICO                 -7.754e-02  1.106e-02  -7.011 2.37e-12 ***
ClubCLUB PLUS                    1.343e-02  1.449e-02   0.927    0.354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 292223  on 210793  degrees of freedom
Residual deviance: 253652  on 210774  degrees of freedom
AIC: 253692

Number of Fisher Scoring iterations: 5
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 40629  8557
         1 13666 13199

               Accuracy : 0.7078
                 95% CI : (0.7045, 0.711)
    No Information Rate : 0.7139
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.3316

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7483
            Specificity : 0.6067
         Pos Pred Value : 0.8260
         Neg Pred Value : 0.4913
             Prevalence : 0.7139
         Detection Rate : 0.5342
   Detection Prevalence : 0.6468
      Balanced Accuracy : 0.6775

       'Positive' Class : 0
```

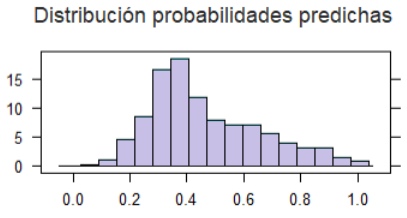


Figura 4: Modelo elegido.

En la Figura 4 se encuentran los resultados obtenidos. Todas las variables resultan significativas, excepto el nivel ClubPlus de la variable **club**.

La tasa de clasificaciones correctas para el conjunto de prueba es de 0.7078. El poder de predicción del modelo baja con respecto al resto de modelos probados, pero es un buen dato ya que en ningún caso se han obtenido valores menores al 70 %. En cambio, la especificidad crece hasta un 60.6 %. Magnífico dato si tenemos en cuenta que al principio teníamos solo un acierto de abandonos de un 24 % (ver sección 7).

En este problema concreto, va a tener mucha importancia el valor de la especificidad y es que es muy importante que este valor sea lo más alto posible.

¿Por qué necesitamos que la especificidad sea bastante alta? Recordemos que intentamos predecir la anulación de un cliente para así poder tomar decisiones a nivel de acciones (marketing) o precio que nos permita retenerlo.

Teniendo esto en cuenta, es preferible asumir fallos en los que se prediga una anulación y que esto sea erróneo, frente a predecir una renovación en clientes que realmente tienen intención de anular. En este último caso, no se le realizarían acciones de retención y acabarían abandonando.

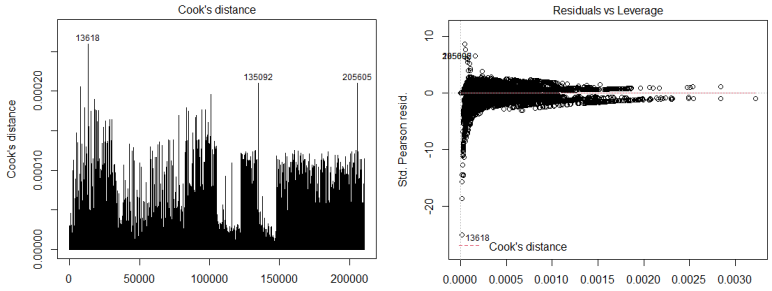
La especificidad mide exactamente esto, qué tan bueno es el acierto de las anulaciones. Cuántos anulados no es capaz de predecir el modelo.

Una vez elegido el modelo, se pasará a estudiar si cumple los supuestos del modelo logístico:

- La variable respuesta es binaria.
- No existe multicolinealidad. Las variables predictoras no están altamente correlacionadas. Se ha usado el factor de inflación de la varianza (VIF) para comprobarlo. Valores del VIF cercanos a 1 indica que no hay correlación entre una variable predictora dada y cualquier otra variable predictora en el modelo.

```
                                GVIF
CreditScoring               1.078844
PrimaVigorTotalDiferencia   1.280907
Perfil                      1.241053
AnyoRenovacionComparativa   1.301376
PaqueteGarantias            1.269403
Club                        1.184644
```

- Hay valores atípicos, pero estos se deben a valores extremos de diferencias de prima. En la gráfica Residuals vs Leverage se observa que aunque sean valores atípicos, no son influyentes.



  Si alguno de los valores fuera influyente habría una línea discontinua roja, que señalaría la distancia de Cook y aparecerían alejados de la misma. Lo que implicaría que tendrían valores de distancia de Cook altos.

- Muestra suficientemente grande.

Con lo que se concluye que se cumplen todos los supuestos del modelo logístico. Ahora la pregunta es, ¿es adecuado este modelo para modelizar los datos? Para dar respuesta a esta pregunta hay que estudiar la bondad del ajuste.

Como se explicó en la parte teórica se disponen de varios test para analizar la bondad del ajuste:

- Se ha comparado el modelo completo con respecto al modelo nulo.

  Como el $p - valor < 0.05$ a un nivel de significación $\alpha = 0.05$ se rechaza la hipótesis nula y se acepta la alternativa en la que se supone que el modelo con los 6 predictores supera al modelo nulo en términos de ajuste.

```
Likelihood ratio test

Model 1: ANULA ~ CreditScoring + PrimaVigorTotalDiferencia + Perfil +
    AnyoRenovacionComparativa + PaqueteGarantias + Club
Model 2: ANULA ~ 1
  #Df  LogLik  Df Chisq Pr(>Chisq)
1  32 -126932
2   1 -146096 -31 38329  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Devianza: en la Figura 4 se puede encontrar el valor de la devianza residual y nula. En este caso hay una reducción de 38329 con una pérdida de 31 grados de libertad (una reducción significativa).

- R^2: se obtiene un valor de 0.1321. Dicho valor no se encuentra en el rango de 0,2 a 0,4 que indica un ajuste muy bueno del modelo. Como tal, el modelo seleccionado probablemente no sea un modelo muy bueno, al menos según esta métrica.

- Wald: la significación de los parámetros mediante el Test de Wald se puede encontrar también en la Figura 4. Se tiene que todos los parámetros son significativos excepto el nivel ClubPlus de la variable **club**.

Para evaluar la capacidad predictiva del modelo calcularemos la curva ROC. Esta curva ofrece mejor resumen de la capacidad del modelo que la tabla de clasificación.

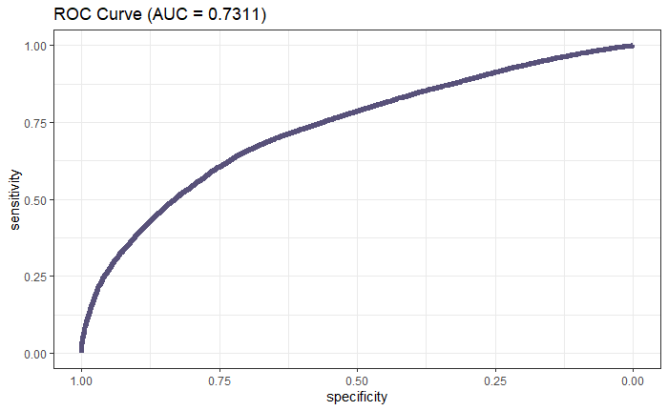


Figura 5: Curva ROC.

Con todo esto, queda claro, que se cumplen los supuestos del modelo logístico, el modelo es adecuado para modelizar los datos y tiene una buena capacidad predictiva.

Antes de pasar a interpretar los parámetros del modelo, quería hacer una última comprobación. Como se comentó al principio, el conjunto de datos elegido contaba con pólizas hasta marzo de 2021 cuya renovación sucedía en marzo de 2022. Mi intención ahora es tomar el mes de abril de 2021, y ver cómo predice la anulación de esos registros el modelo seleccionado.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2895  560
         1 1006 1277

               Accuracy : 0.7271
                 95% CI : (0.7154, 0.7386)
    No Information Rate : 0.6799
    P-Value [Acc > NIR] : 4.169e-15

                  Kappa : 0.4109

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7421
            Specificity : 0.6952
         Pos Pred Value : 0.8379
         Neg Pred Value : 0.5594
             Prevalence : 0.6799
         Detection Rate : 0.5045
   Detection Prevalence : 0.6021
      Balanced Accuracy : 0.7186

       'Positive' Class : 0
```

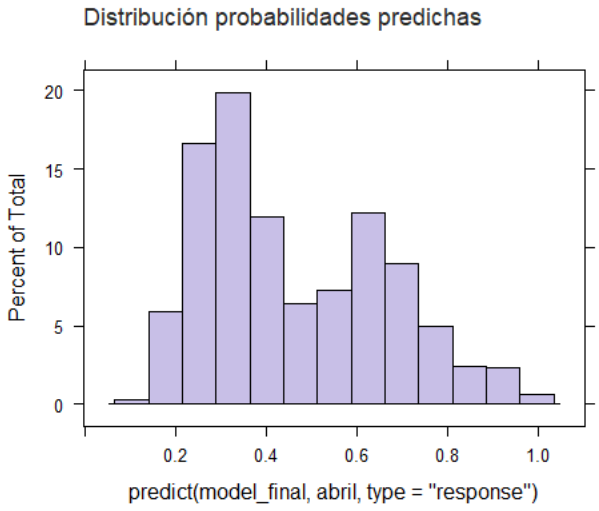


Figura 6: Resultado mes de abril.

Se obtiene un porcentaje de acierto de un 72.7 % y una especificidad de un 70 %. Un resultado bastante bueno y que refuerza una vez mas el modelo seleccionado.

Una posible aplicación a nivel de negocio que se le puede dar a la probabilidad devuelta por el modelo, es la de generar un conjunto de registros ”indecisos”, es decir, fijar un umbral superior para asumir que a partir de ese punto de corte, es bastante probable que la persona anule, y un umbral inferior a partir del cual, es muy probable que la persona renueve. Esto daría lugar a un tercer grupo de ”indecisos”, sobre el cual se podrían aplicar acciones.

Si se hace esto para el conjunto de abril, fijándose los umbrales en 0.3 y 0.7, se obtendría la siguiente división.

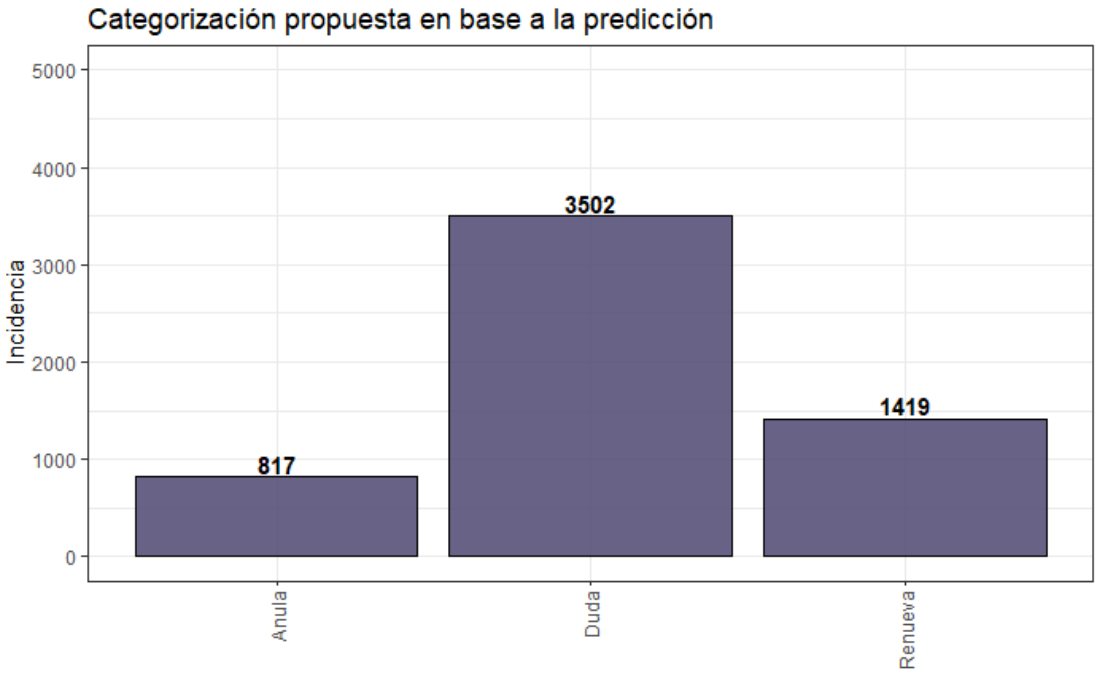


Figura 7: Categorización de la anulación de abril.

Y ahora sí, se pasará a interpretar los parámetros del modelo, lo cual aportará información interesante para el negocio y además permitirá responder a una importante pregunta: ¿Cómo afecta la subida de X€ en la decisión de anular/renovar?

El ajuste del modelo lleva a la siguiente interpretación de los parámetros

(ver Figura 8):

- $\theta_{diferenciadeprima} = e^{0.006037} = 1.006$. La ventaja de anular frente a renovar se multiplica por 1.006 cuando aumenta en una unidad la diferencia de prima, manteniéndose el resto de variables constantes.

- Si se quiere hacer una subida de 5€ en el importe total de cada póliza, manteniéndose el resto de variables constante, la ventaja de anular frente a renovar se verá multiplicada por $1.006^5 = 1.03$. Y su intervalo de confianza será $(1.02995, 1.031616)$.

- $\theta_{Perfil1RECARGO,Perfil2RECARGOS} = e^{0.7341} = 2.084$. La ventaja de anular frente a renovar se multiplica por 2.084 cuando se pasa de un perfil de 2 Recargos a 1 Recargo.

- $\theta_{PerfilBASE,Perfil2RECARGOS} = e^{0.8206} = 2.272$. La ventaja de anular frente a renovar se multiplica por 2.272 cuando se pasa de un perfil de 2 Recargos a Base.

- $\theta_{PerfilBONIFICADO,Perfil2RECARGOS} = e^{0.7385} = 2.093$. La ventaja de anular frente a renovar se multiplica por 2.093 cuando se pasa de un perfil de 2 Recargos a Bonificado.

- $\theta_{Añoderenovación} = e^{-0.1107} = 0.8952$. La ventaja de anular frente a renovar se multiplica por 0.8952 cuando aumenta en una unidad la antigüedad de la póliza.

- La ventaja de anular frente a renovar se multiplica por 0.5146 cuando se pasa de primera renovación a 6 renovación. Es decir, una póliza que se enfrente a su sexta renovación tiene la mitad de probabilidades de anular que una de primera renovación.

- $\theta_{CreditScoring3,CreditScoring6} = e^{-0.4385} = 1.55$. La ventaja de anular frente a renovar se multiplica por 1.55 cuando se pasa de un nivel crediticio máximo a un nivel crediticio nivel "3" . Es decir, la probabilidad de anular se ve aumentada un 50 %.

- $\theta_{PaqueteGarantiasTODORIESGO,PaqueteGarantiasROBOEINCENDIO} = \frac{\theta_{PaqueteGarantiasTODORIESGO,PaqueteGarantiasTERCEROS}}{\theta_{PaqueteGarantiasROBOEINCENDIO,PaqueteGarantiasTERCEROS}} = \frac{0.698}{0.7163} = 0.97$. La ventaja de anular frente a renovar se multiplica por 0.97 cuando se pasa de un todo riesgo a un robo+incendio.

	OR	cociente ventajas	2.5 %	97.5 %
(Intercept)	-0.9838	0.3739	0.3444	0.4059
CreditScoring5	0.1291	1.138	1.097	1.18
CreditScoring4	0.2997	1.349	1.304	1.397
CreditScoring3	0.4385	1.55	1.492	1.611
CreditScoring1 - 2	0.9739	2.648	2.474	2.835
CreditScoringA	0.9432	2.568	2.35	2.807
CreditScoringB	1.019	2.771	2.522	3.045
CreditScoringC	0.9799	2.664	2.395	2.964
CreditScoringD	1.4	4.055	3.518	4.674
CreditScoringF - E	1.421	4.142	3.505	4.895
CreditScoring-	1.724	5.605	5.369	5.851
PrimaVigorTotalDiferencia	0.006037	1.006	1.006	1.006
Perfil1 RECARGO	0.7341	2.084	1.925	2.256
PerfilBASE	0.8206	2.272	2.104	2.453
PerfilBONIFICADO	0.7385	2.093	1.945	2.252
AnyoRenovacionComparativa	-0.1107	0.8952	0.8916	0.8989
PaqueteGarantiasROBO E INCENDIO	-0.3337	0.7163	0.6995	0.7334
PaqueteGarantiasTODO RIESGO	-0.3595	0.698	0.6793	0.7173
ClubCLUB BASICO	-0.093	0.9112	0.8917	0.9312
ClubCLUB PLUS	-0.02205	0.9782	0.9508	1.006

Figura 8: Coeficientes, cociente de ventajas e intervalos de confianza.

7 Otros aspectos prácticos

En esta sección se detallarán todos los pasos realizados hasta llegar al modelo final. Desde el análisis de las variables hasta las distintas pruebas realizadas con diferentes modelos.

Como se ha comentado en la sección 5, el conjunto de datos tiene originalmente un total de 44 variables. Son muchas variables y no se conoce si todas van a aportar algún valor, por lo que en primer lugar se va a realizar un análisis descriptivo y algunas representaciones gráficas que nos permitan hacer un primer descarte.

Aquí mostraré algunos de los resultados más llamativos o que me hayan llevado a hacerme algunas preguntas.

En primer lugar empezaré hablando de las variables que hacen referencia a los **regalos** realizados a los clientes. Estas variables pueden tomar el valor 1, si el cliente a recibido dicho regalo durante el año en cuestión, o 0 en caso contrario.

Estas variables me parecen muy interesantes de analizar pero desgraciadamente a penas se han realizado regalos (ver Figura 9), por lo que el volumen entre ambos grupos (recibe dicho regalo o no lo recibe), es extremadamente diferente. Teniendo esto en cuenta se ha decidido descartarlas del estudio.

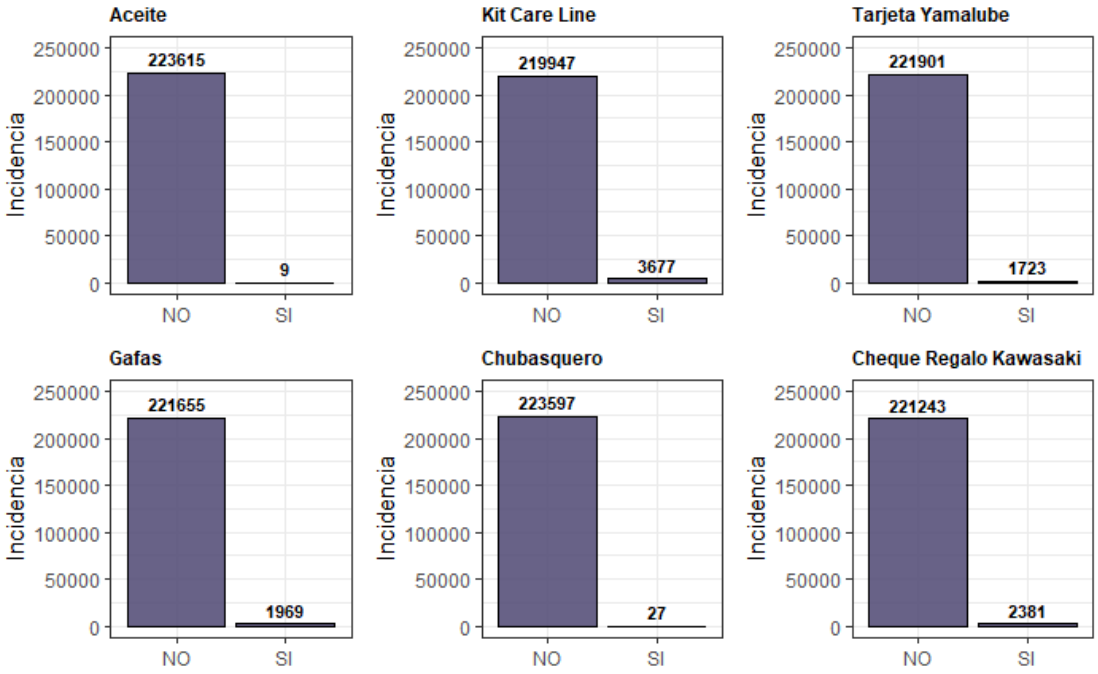


Figura 9: Distribución por tipo de regalo.

Lo mismo sucede con las variables que hacen referencia al **uso de una cobertura** en concreto. Para poder interpretar los gráficos se necesita saber que el 9 denota no tener contratada la cobertura, el 0 que la tiene pero no ha hecho uso y el 1 que sí ha hecho uso.

Se observa rápidamente que el número de usos es ínfimo (ver Figura 10). Por lo que estas variables también serán descartadas.

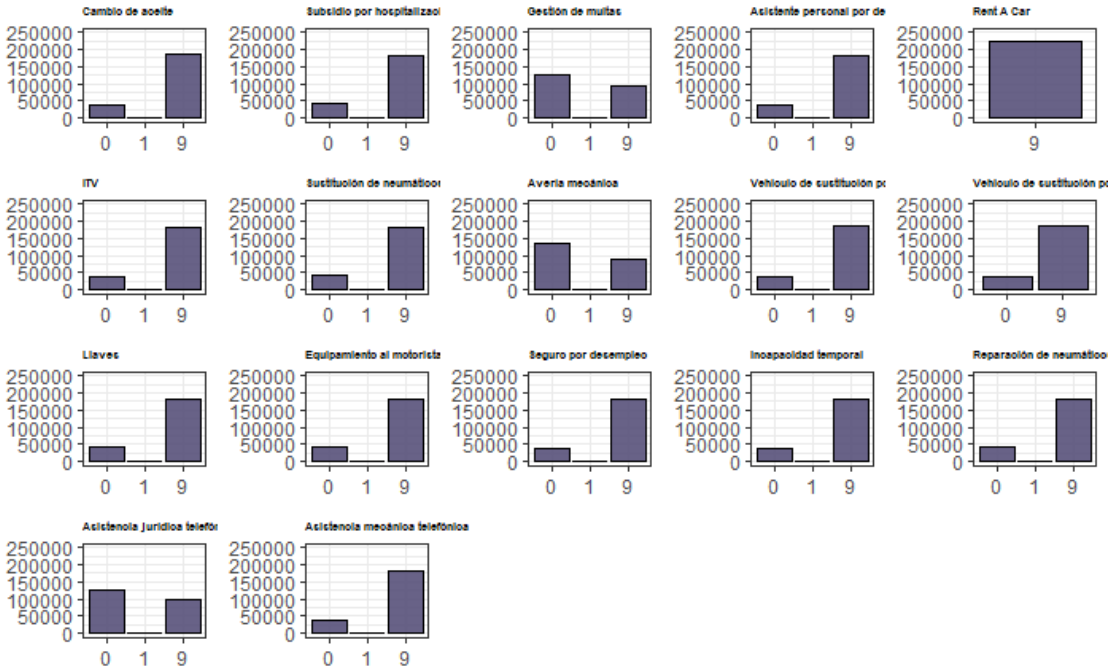


Figura 10: Distribución por tipo de cobertura.

Otras variables importantes son la **edad del tomador**, **años de carnet** y **años del vehículo**. En la Figura 11 encontraremos una visualización de estas variables.

Se trata de un gráfico de doble eje. En el eje primario se mide el número de elementos de cada grupo. En el eje secundario se mide el porcentaje de anulación asociado a dicho grupo.

Se propone categorizar estas variables agrupando por tramos de edad o años (dependiendo de la variable) con valores de anulación similares.

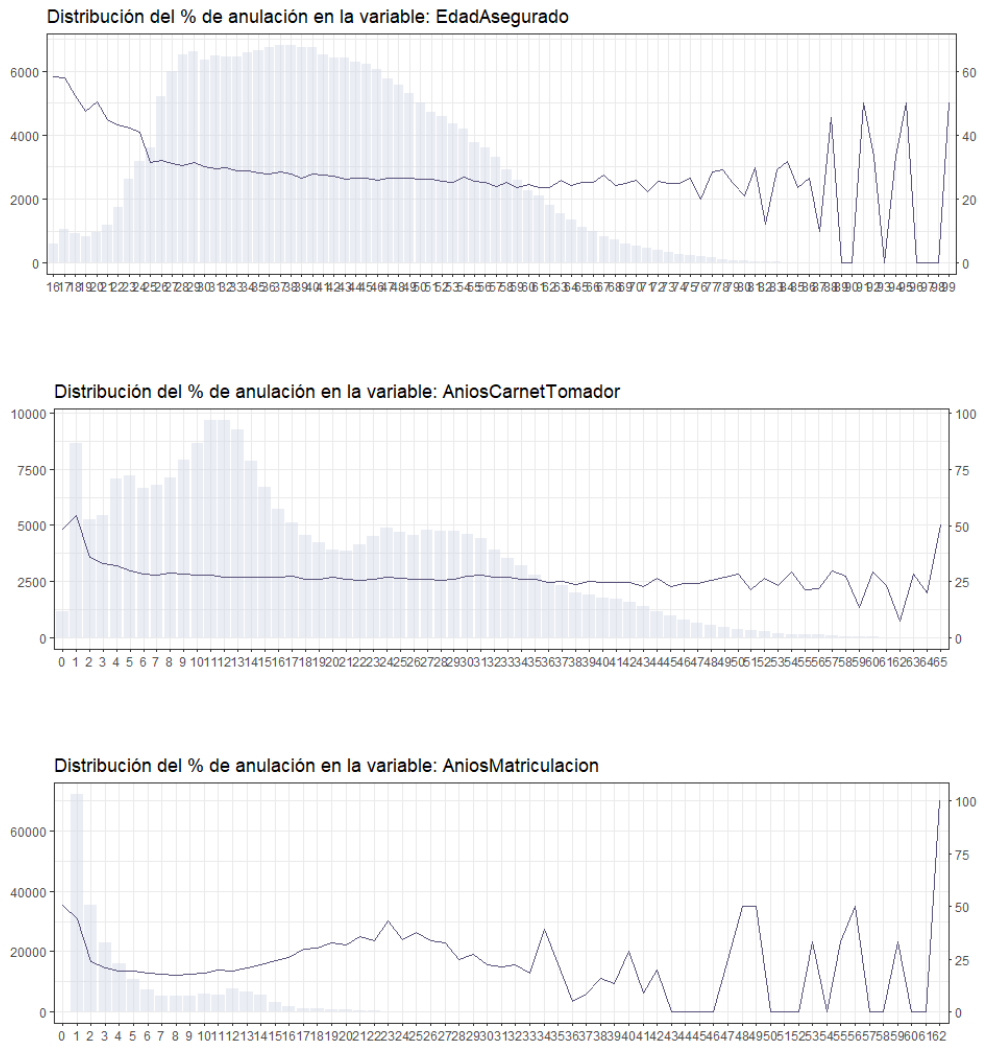


Figura 11: Distribución por tipo de cobertura.

Los grupos generados para cada una de las variables y sus respectivos porcentajes de anulación se pueden consultar en la Figura 12. Con la agrupación sugerida se recoge con más claridad la evolución del porcentaje de anulación a medida que aumenta la edad o los años de carnet/matriculación.

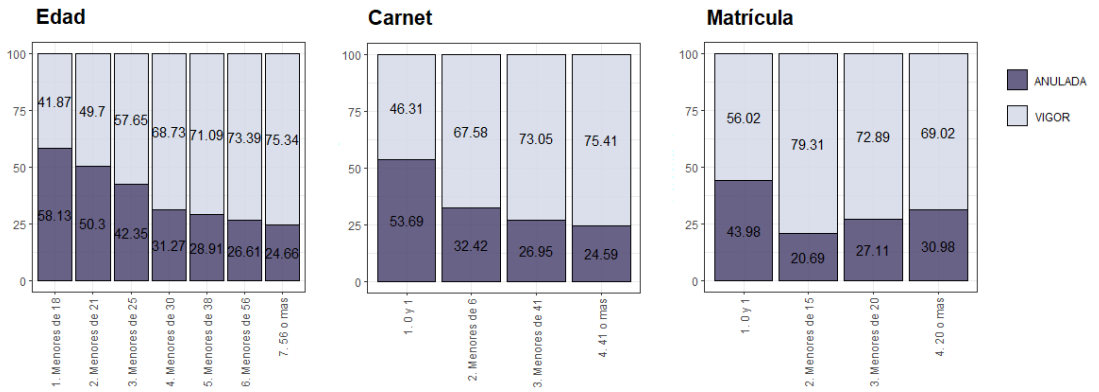
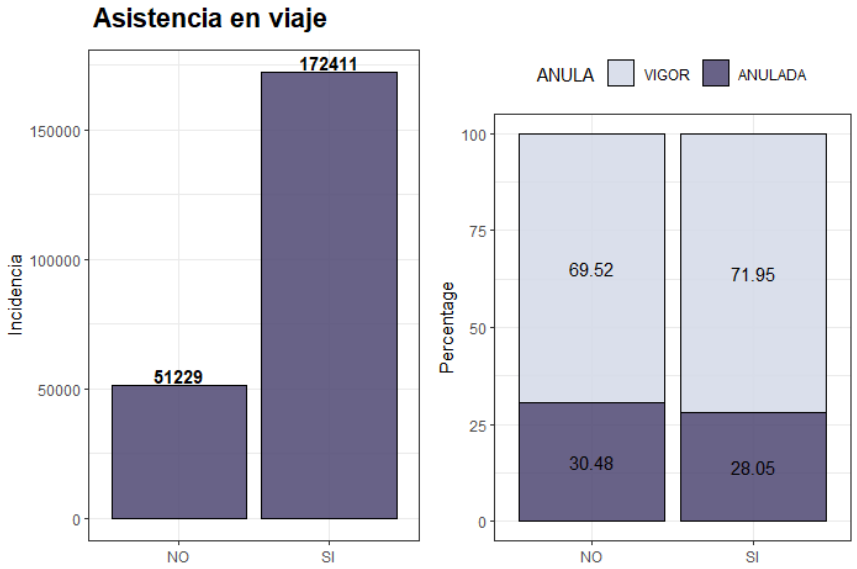


Figura 12: % de anulación en la edad del tomador, años de carnet y años de la moto.

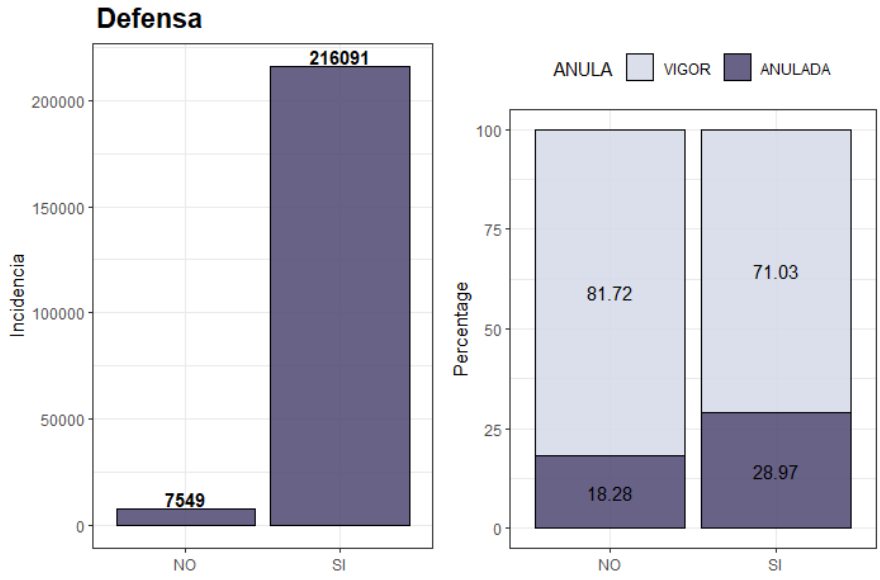
Por otro lado, tenemos 3 variables que hacen referencia a extras contratados aparte del tipo de paquete (tercero, todo riesgo, etc). Estas son la **asistencia en viaje**, **defensa** y **accidentes**.

- Asistencia en viaje: Ambos grupos están bien representados pero muestran una anulación similar. Se hará una prueba Chi-Cuadrado para ver si existe relación entre dicha variable y la variable respuesta.



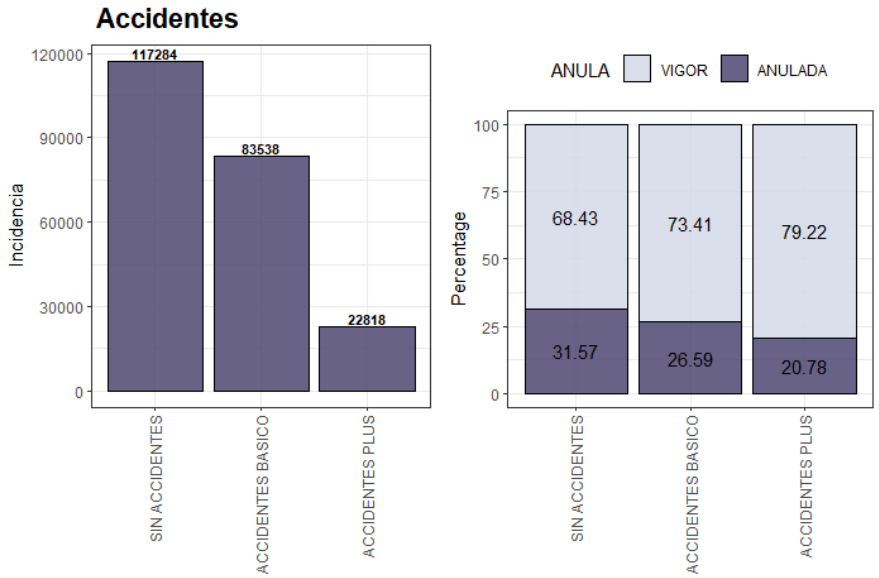
Al realizar el contraste de hipótesis se obtiene un p-valor < 0.05, por el cuál se determina que existe relación entre ambas variables y por tanto, dicha variable será propuesta para el modelo.

- Defensa: uno de los grupos tiene poco volumen de datos. Se hará una prueba Chi-Cuadrado para ver si existe relación entre dicha variable y la variable respuesta.



Al realizar el contraste de hipótesis se obtiene un p-valor < 0.05, por el cuál se determina que existe relación entre ambas variables y por tanto, dicha variable será propuesta para el modelo.

- Accidentes: se diferencias dos tipos, la básica y la plus. Todos los grupos están bien representados y además existen diferencias en la renovación en cada uno de ellos.



También se dispone de datos de localización geográfica, como es la **provincia**. Debido al gran número de provincias que se pueden encontrar, se ha decidido hacer una agrupación en función del porcentaje de anulación de cada una.

Cód. Provincia	% Anulación
08	22'68%
27	24'55%
07	25'10%
51	25'76%
17	26'39%
15	27'33%
29	27'48%
28	27'67%
31	28'29%
40	28'66%
33	30'01%
43	30'23%
20	30'28%

Cód. Provincia	% Anulación
48	30'40%
46	30'71%
41	30'90%
36	31'06%
01	31'63%
25	32'05%
42	32'46%
44	32'48%
52	32'88%
12	32'89%
19	33'33%
49	33'33%
38	33'46%

Cód. Provincia	% Anulación
22	33'52%
32	33'71%
26	34'05%
50	34'10%
35	34'31%
47	34'42%
37	34'67%
18	34'74%
24	34'89%
2 1	35'67%
03	35'69%
11	35'75%
04	35'87%

Cód. Provincia	% Anulación
39	35'92%
09	36'28%
45	36'88%
13	37'03%
16	38'20%
05	38'25%
34	39'40%
02	40'39%
30	40'42%
14	41'54%
10	41'60%
06	45'78%
23	49'17%

Figura 13: % de anulación por provincias.

Como se puede observar en la Figura 13, hay bastantes diferencias dependiendo de la provincia en la que nos fijemos. Se puede encontrar desde un

22 % de anulación en Barcelona a casi un 50 % en Jaén.

Se sugiere una categorización en función de la anulación, agrupando aquellas provincias con un comportamiento similar. Como resultado se obtienen 4 zonas:

- Zona 1: Baleares, Barcelona, Lugo y Ceuta.
- Zona 2: La Coruña, Girona, Guipúzcoa, Madrid, Málaga, Navarra, Asturias, Segovia, Sevilla, Tarragona, Valencia y Vizcaya.
- Zona 3: Álava, Alicante, Almería, Cádiz, Castellón, Granada, Guadalajara, Huelva, Huesca, León, Lérida, La Rioja, Orense, Las Palmas, Pontevedra, Salamanca, Santa Cruz de Tenerife, Cantabria, Soria, Teruel, Valladolid, Zamora, Zaragoza y Melilla.
- Zona 4: Albacete, Ávila, Badajoz, Burgos, Cáceres, Ciudad Real, Córdoba, Cuenca, Jaén, Murcia, Palencia y Toledo.

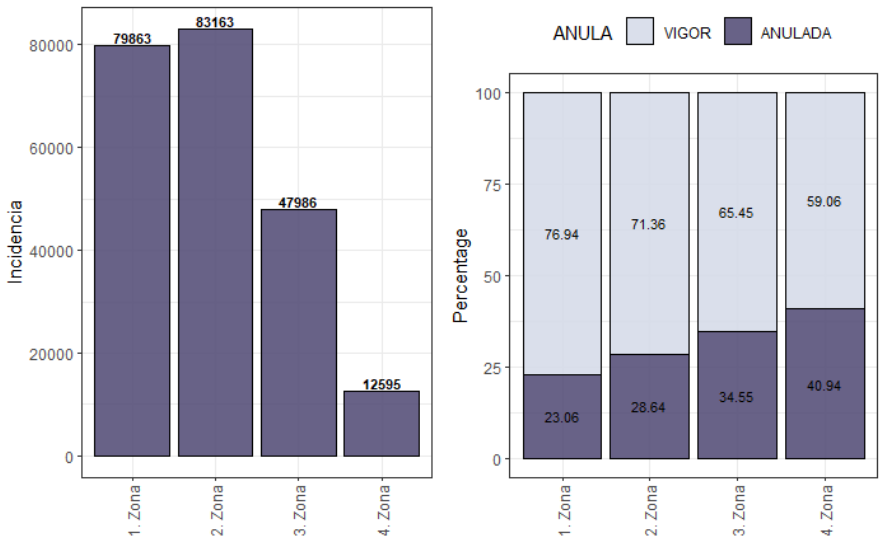


Figura 14: % de anulación por zonas.

El test Chi-Cuadrado aplicado a esta variable frente a la variable respuesta reporta un p-valor < 0.05, por el cuál se determina que existe relación entre

ambas variables y por tanto, dicha variable podría ser propuesta para el modelo. Aún así no se tendrá en cuenta para generar el modelo ya que dicha agrupación no tiene lógica.

Se estaría agrupando en un mismo grupo a provincias totalmente desiguales como pueden ser Barcelona y Ceuta. Por lo que se debería hacer un análisis más profundo aplicando por ejemplo técnicas de clustering para poder hacer una agrupación más lógica.

Por otro lado tenemos las variables **marca** y **división**. Ambas variables son muy semejantes. La marca hace referencia a la marca del vehículo, y la división son agrupaciones de agentes (entendiéndose agente por quién hace la venta). En la empresa se aplican reglas de precios según la división, pero en una misma división se puede encontrar diferentes marcas, como puede ocurrir, por ejemplo en la división de **Ventas Directas**.

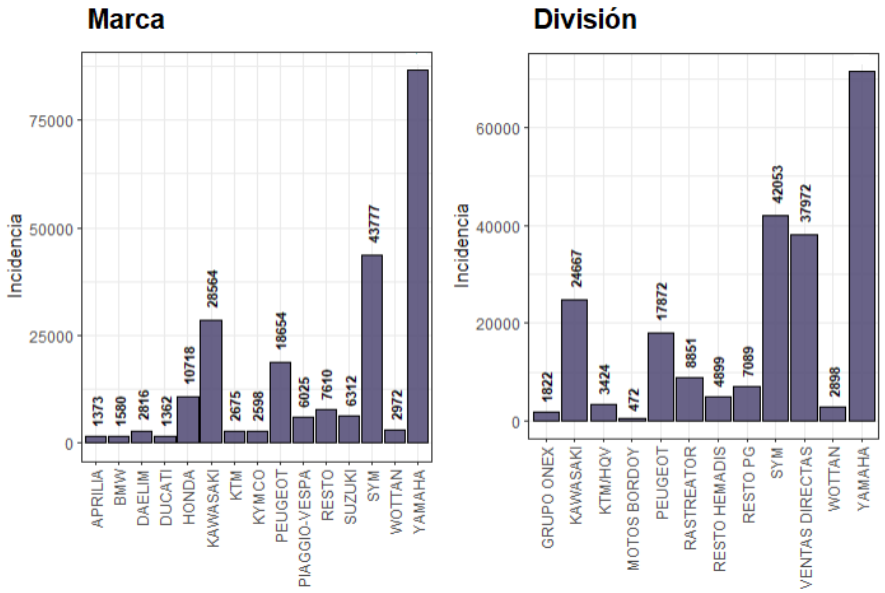


Figura 15: Distribución marca y división.

Para ambas variables se han agrupado las marcas/divisiones menos relevantes en un grupo llamado ”resto”. La forma de proceder en cada variable

ha sido la siguiente: para las marcas, se han distinguido aquellas que tienen un mayor volumen y para la división se han dejado los negocios más importantes, para los cuáles se aplican reglas de negocio independientemente de su volumen, como ocurre en el caso del grupo "Bordoy".

El hecho de que la división afecte en el precio, hace pensar que esta variable debe de ser propuesta para el modelo. En este punto cabría preguntarse si sería necesario incluir la marca del vehículo también, ya que son bastante similares. Basándonos en que se desconoce como actúan el resto de aseguradoras, si aplican las reglas directamente sobre la marca o no, se propone utilizar también esta variable en el modelo.

Una de las variables que no ha mostrado relación con la variable respuesta en el test Chi-Cuadrado es **Grupo Cartera**. Esta variable diferencia unas pólizas de otras según si pertenecen al grupo de control o no. A las pólizas del grupo de control no se le aplican reglas de negocio.

En la Figura 16 queda claro que no existe una diferencia significativa entre ambos grupos.

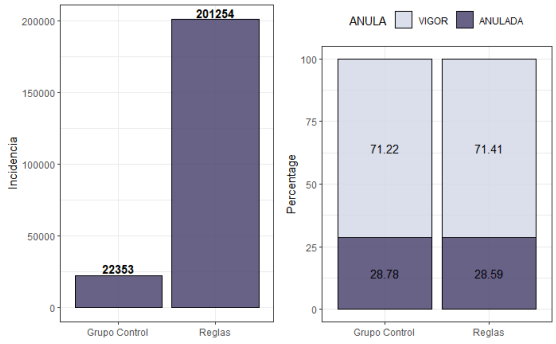


Figura 16: % de anulación grupo cartera.

A continuación se mostrarán el resto de variables que se proponen para el modelo. Para estas no se ha tenido que realizar ninguna categorización ni comprobación, ni razonamiento para discernir entre unas y otras. Aun así se comentarán de forma muy resumida.

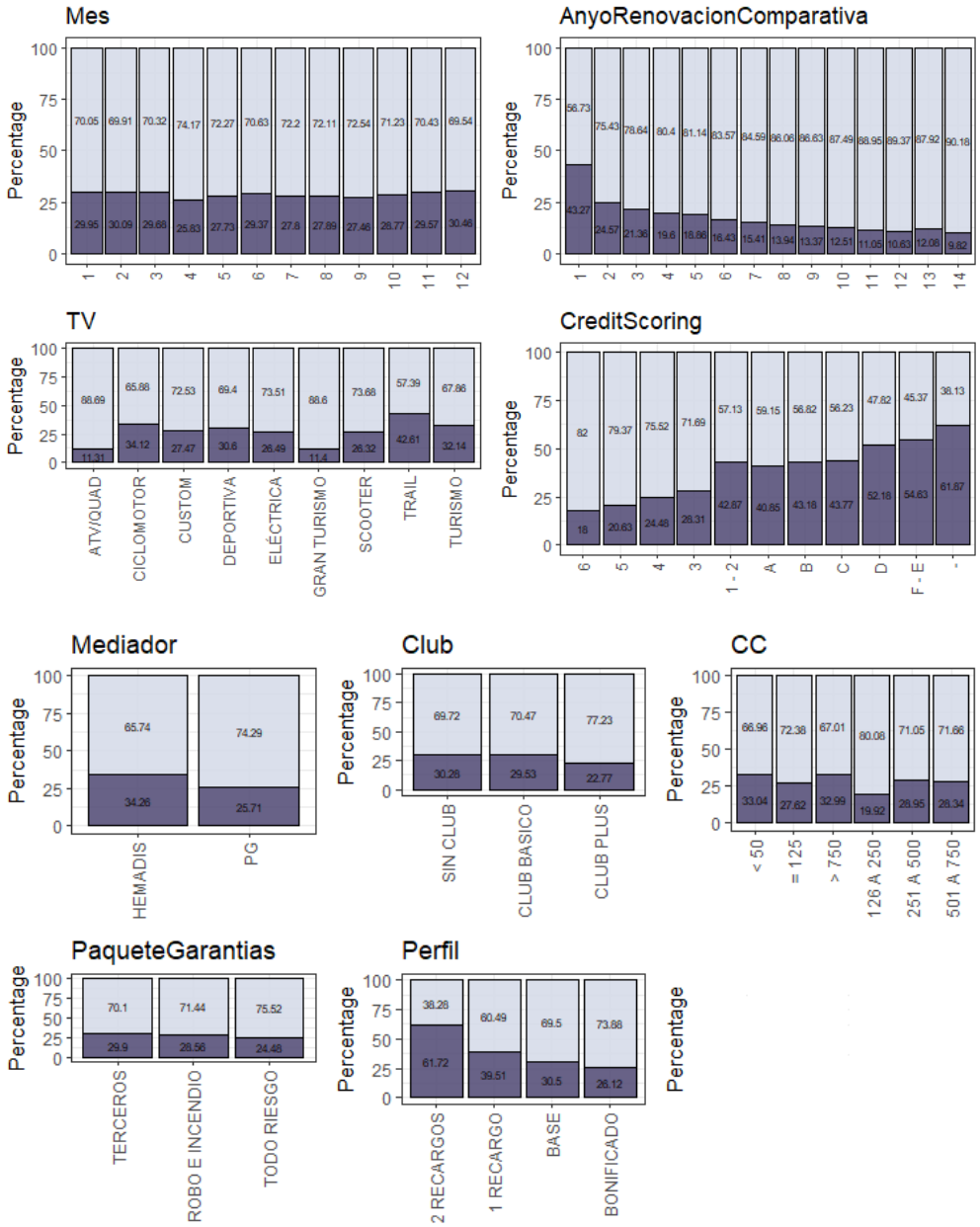


Figura 17: % de anulación del resto de variables.

- Las variables **año de renovación**, **nivel crediticio** y **perfil** son cualitativas y con un cierto orden. Muestran una perfecta evolución del porcentaje de anulación a medida que se crece en dicho ”orden”. Son variables muy importantes para la empresa.

- No existen grandes diferencias entre unos **meses** y otros. Si se observa menor anulación en los meses de verano con respecto a los de invierno.
- Gran diferencia en el porcentaje de anulación según el **mediador** del seguro.
- A pesar de que quienes contratan el **club** plus, tienen una prima superior, muestran una anulación 8 puntos inferior con respecto a los que no tienen contratado el club.

Por último, pero no menos importante, están las variables que hacen referencia a la prima del seguro.

Por un lado, se considerará la variable **prima total** la cuál mide la cantidad total a pagar por el cliente en el momento de la renovación. Y por otro lado estará la variable **diferencia de prima** que indica la diferencia de prima con el año anterior.

A continuación se puede observar una pequeño análisis descriptivo de ambas variables:

```
PrimaVigorTotal
       n  missing distinct     Info     Mean      Gmd      .05      .10
  227336        0    33334    0.999    166.2    116.5      0.0      0.0
     .25      .50      .75      .90      .95
   104.8    157.3    226.9    294.2    348.9

lowest :   -8.09   -0.01    0.00    0.51    1.12
highest: 1304.09 1313.66 1319.48 1341.42 1491.59
------------------------------------------------------------------------
PrimaVigorTotalDiferencia
       n  missing distinct     Info     Mean      Gmd      .05      .10
  227297       39    33234        1     42.4    83.37   -22.49   -12.01
     .25      .50      .75      .90      .95
   -4.02     5.07    72.96   144.27   189.51

lowest : -962.31 -891.09 -816.16 -801.95 -758.59
highest: 2009.60 2322.98 2672.49 2853.81 4797.77
```

Figura 18: Descriptivos de las variables prima total y diferencia de prima.

Llegados a este punto en el que hemos estudiado todas las variables que se van a proponer para el modelo, faltaría comentar la variable respuesta. Como se mencionó en la sección 5, el objetivo de la parte práctica de este trabajo se centra en predecir si al vencimiento de la póliza, el tomador de la póliza va a anular o renovar. Por lo que usaremos la variable **estado de la póliza** como variable respusta del modelo. Esta variable tomará el valor 1 si anula y 0 si renueva.

Como se observa en la Figura 19, estamos ante un conjunto de datos en el que la variable respuesta no está balanceada. Encontramos una relación 60-40. Esto tendrá que tenerse en cuenta a la hora de aplicar el modelo.

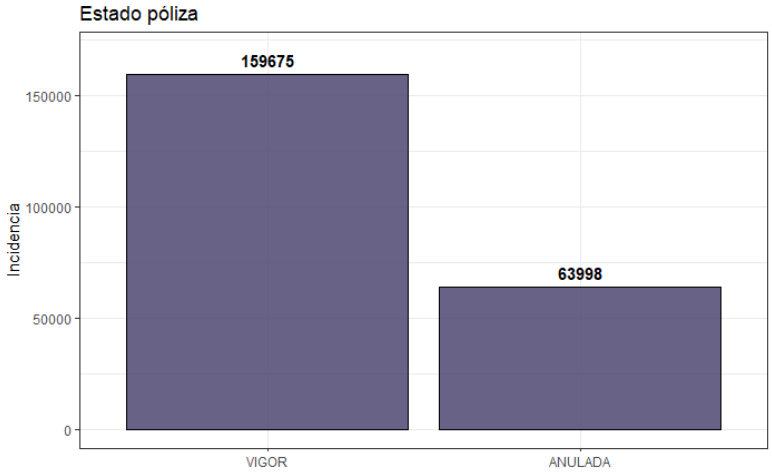


Figura 19: Distribución variable respuesta.

Ya se ha hecho una revisión de todas las variables que se van a usar, tanto las predictoras como la variable respuesta, por lo que ahora se pasará a hacer una revisión de las correlaciones entre las variables haciendo uso de los coeficientes de correlación de Pearson y Spearman.

Como se mencionó en la parte teórica, para las variables cualitativas se usará el coeficiente de correlación de Spearman.

En la Figura 20 se observa lo siguiente:

- Una correlación positiva en el tándem Perfil, fr_edad, fr_carnet y fr_matricula. Como se explicó anteriormente el perfil es una variable calculada entre la edad y los años de carnet del tomador de la póliza.
- Marca y División, como ya comentamos, también muestran una alta correlación positiva.
- Mediador y club van de la mano porque la gran mayoría de los vehículos de uno de los dos mediadores que se disponen, no tienen club.

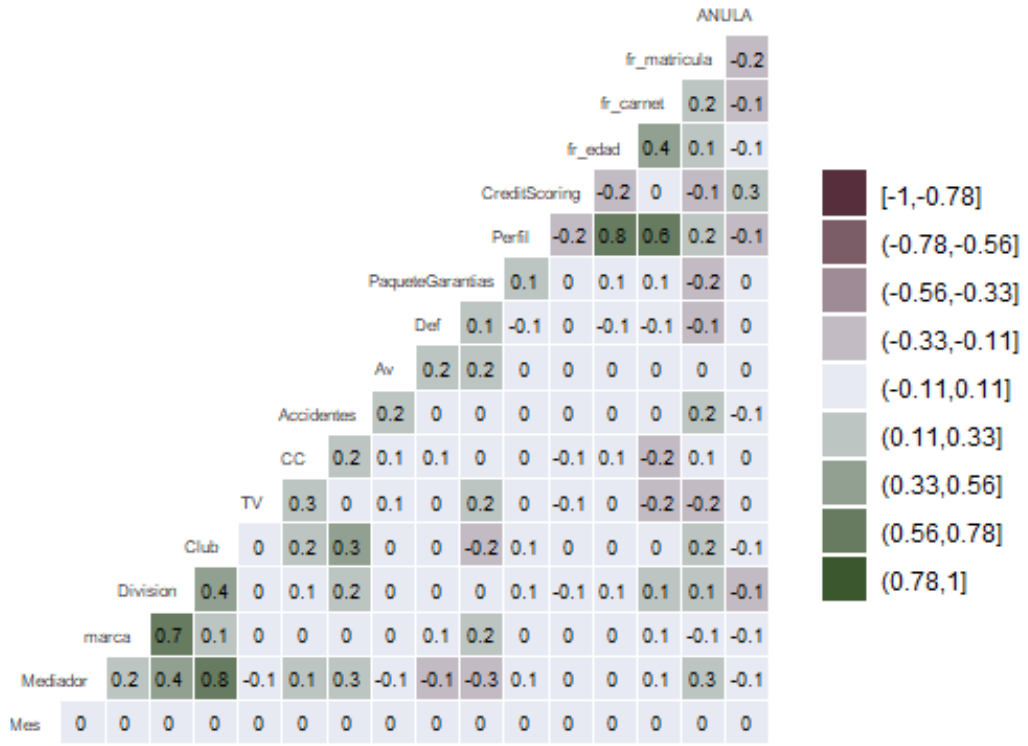


Figura 20: Correlación variables cualitativas.

En cambio para las variables cuantitativas se usará el coeficiente de correlación de Pearson. Se obtiene una correlación positiva entre **diferencia de prima** y la variable respuesta. Todo lo contrario ocurre con la variable **año de renovación**.

Figura 21: Correlación variables cuantitativas.

Ahora sí, ha llegado el momento de ajustar el modelo. Para ello se va dividir el conjunto de datos en dos subconjuntos, uno de entrenamiento y otro de prueba de forma que el primero conserve el 70 % de los registros y el segundo, el 30 % restante.

Es importante que esta división se haga de forma aleatoria para así asegurar la máxima variabilidad de los datos en ambos subconjuntos. Además como nuestro conjunto de datos no está balanceado con respecto a la variable respuesta, se debe comprobar que el conjunto de entrenamiento resultante cumpla la misma relación, o de lo contrario los resultados predichos por el modelo estarán sesgados.

En este caso concreto, en el que recordemos, los datos se extrajeron con una consulta a base de datos, se tiene que tener más cuidado aún a la hora de dividirlo, ya que dicha consulta estaba ordenada por años. Esto puede provocar que haya más valores anulados concentrados al principio o al final del conjunto de datos original, dependiente del número de anulaciones que haya habido cada año.

Al principio se cayó en el error de no hacer el reparto de manera aleatoria lo que provocó que el conjunto de entrenamiento tuviera una relación 74-24 para la variable respuesta, que a su vez repercutió en el poder de predicción de los modelos que se van a presentar a continuación. Obteniéndose en todos los casos peores predicciones.

Con todo lo comentado anteriormente, ya se podría ajustar el modelo. Se va a proceder de dos formas, una tanteando y otra haciendo uso de stepwise.

Para empezar con el tanteo se eligieron variables muy importantes a nivel de negocio como son el **perfil** y el **nivel crediticio**. Además, estas variables mostraron una buena diferenciación en cuanto al porcentaje de anulación en cada uno de sus niveles. También pareció interesante considerar la variable de **diferencia de prima** porque tiene un impacto directo sobre el cliente. Se quiso comprobar el poder predictivo de estas tres variables de manera independiente.

Los resultados obtenidos por los tres modelos de regresión logística binaria simple fueron:

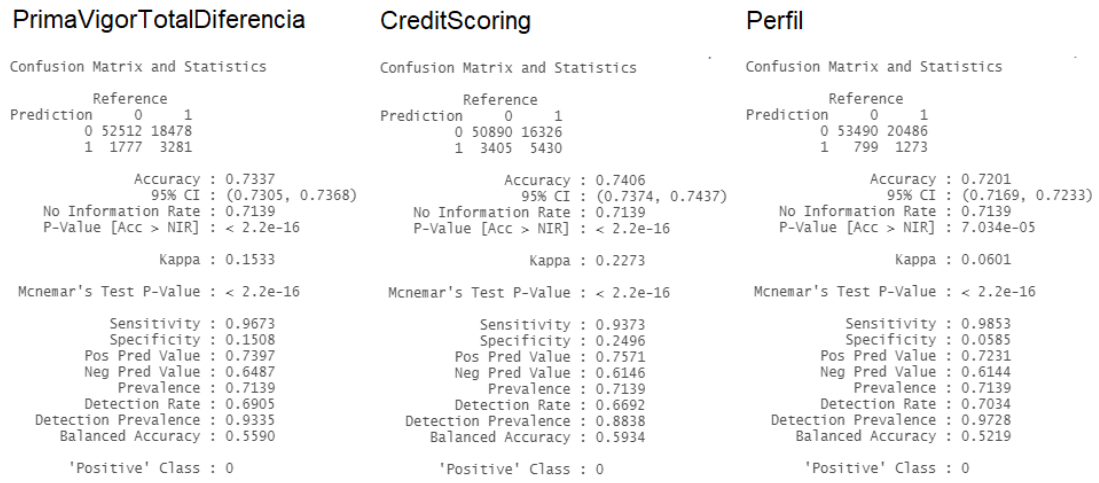


Figura 22: Modelos logísticos simples.

Se observa como el que proporciona mayor porcentaje de acierto es el modelo que tiene como única variable predictora la que hace referencia al nivel crediticio del tomador del seguro. Todos tienen valores similares de sensibilidad pero valores muy dispares de especificidad.

Teniendo esto en cuenta, de los tres modelos se tomará el segundo, el que hace referencia al nivel crediticio. Ahora se añadirá a este modelo una segunda variable. Probaremos con el perfil y la diferencia de prima.

CreditScoring y Perfil

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 50805 16216
         1  3490  5540

               Accuracy : 0.7409
                 95% CI : (0.7378, 0.744)
    No Information Rate : 0.7139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2308

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9357
            Specificity : 0.2546
         Pos Pred Value : 0.7580
         Neg Pred Value : 0.6135
             Prevalence : 0.7139
         Detection Rate : 0.6680
   Detection Prevalence : 0.8813
      Balanced Accuracy : 0.5952

       'Positive' Class : 0
```

CreditScoring y PrimaVigorTotalDiferencia

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 51137 15604
         1  3158  6152

               Accuracy : 0.7533
                 95% CI : (0.7502, 0.7564)
    No Information Rate : 0.7139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2711

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9418
            Specificity : 0.2828
         Pos Pred Value : 0.7662
         Neg Pred Value : 0.6608
             Prevalence : 0.7139
         Detection Rate : 0.6724
   Detection Prevalence : 0.8776
      Balanced Accuracy : 0.6123

       'Positive' Class : 0
```

Figura 23: Modelos logísticos con 2 variables predictoras.

De entre los dos modelos propuestos, el segundo mejora tanto el porcentaje de acierto, en 1.27 puntos porcentuales, como la especificidad en 3.32 puntos porcentuales. Además, mejoran el AIC y el BIC, pasando de 165800 a 155802, en el primero de los casos, y de 165908.5 a 155921.2 en el segundo.

Con los modelos ajustados hasta el momento se puede tener una idea del nivel de predicción que podemos obtener, aunque sea mínimamente, ya que no se van a mostrar el resultado de hacer todas las combinaciones posibles entre todas las variables. Por ese motivo, se va a dar paso a la aplicación del método stepwise el cuál se aplicará en ambas direcciones.

Los resultados obtenidos fueron:

```
Step:  AIC=150611.1
ANULA ~ PrimaVigorTotalDiferencia + CreditScoring + AnyoRenovacionComparativa +
    fr_provincia + Division + fr_edad + PrimaVigorTotal + PaqueteGarantias +
    Av + Perfil + CC + fr_matricula + Mes + TV + Club + Accidentes +
    fr_carnet + marca

                           Df Deviance    AIC
<none>                          150411 150611
+ Mediador                  1   150411 150613
+ Def                       1   150411 150613
- marca                    14   150448 150620
- fr_carnet                 3   150429 150623
- Accidentes                2   150438 150634
- TV                        8   150460 150644
- Club                      2   150452 150648
- Perfil                    3   150460 150654
- Mes                      11   150518 150696
- fr_matricula              3   150518 150712
- Division                 11   150543 150721
- fr_edad                   6   150557 150745
- Av                        1   150560 150758
- CC                        5   150569 150759
- PaqueteGarantias          2   150998 151194
- AnyoRenovacionComparativa 13   151163 151337
- PrimaVigorTotal           1   151283 151481
- fr_provincia              3   151376 151570
- PrimaVigorTotalDiferencia 1   154697 154895
- CreditScoring            10   157112 157292
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 51078 14668
         1  3211  7091

               Accuracy : 0.7649
                 95% CI : (0.7619, 0.7679)
    No Information Rate : 0.7139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3167

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9409
            Specificity : 0.3259
         Pos Pred Value : 0.7769
         Neg Pred Value : 0.6883
             Prevalence : 0.7139
         Detection Rate : 0.6717
   Detection Prevalence : 0.8645
      Balanced Accuracy : 0.6334

       'Positive' Class : 0
```

Figura 24: Modelo stepwise.

De las 20 variables propuestas, el modelo selecciona 18 y deja fuera las variables **mediador** y **defensa**. Se obtiene un porcentaje de acierto de un 76.5 % y la especificidad aumenta 4.3 puntos porcentuales.

Como se puede observar, a penas hay diferencias entre el modelo sencillo que se ha planteado anteriormente y el modelo seleccionado por el método stepwise. Por el principio de parsimonia, ante dos modelos que explican lo mismo, se elige el más sencillo.

Por lo que de aquí en adelante continuaremos tomando como base el modelo que considera las variables **nivel crediticio** y **diferencia de prima**. Ahora se va a intentar mejorar este modelo aplicando alguna técnica de las mencionadas en la teoría para balancear el conjunto de datos.

Las técnicas que se van a aplicar sobre el conjunto de entrenamiento son:

- Upsampling: este método aumenta el tamaño de la clase minoritaria mediante el muestreo con reemplazo para que las clases tengan el mismo tamaño.

- Downsampling: en contraste con el método anterior, este reduce el tamaño de la clase mayoritaria para que sea igual o más cercano al tamaño de la clase minoritaria simplemente tomando una muestra aleatoria.
- ROSE (Random oversampling examples): es un método híbrido que trata de reducir la muestra de la clase mayoritaria y crear nuevos puntos artificiales en la clase minoritaria.

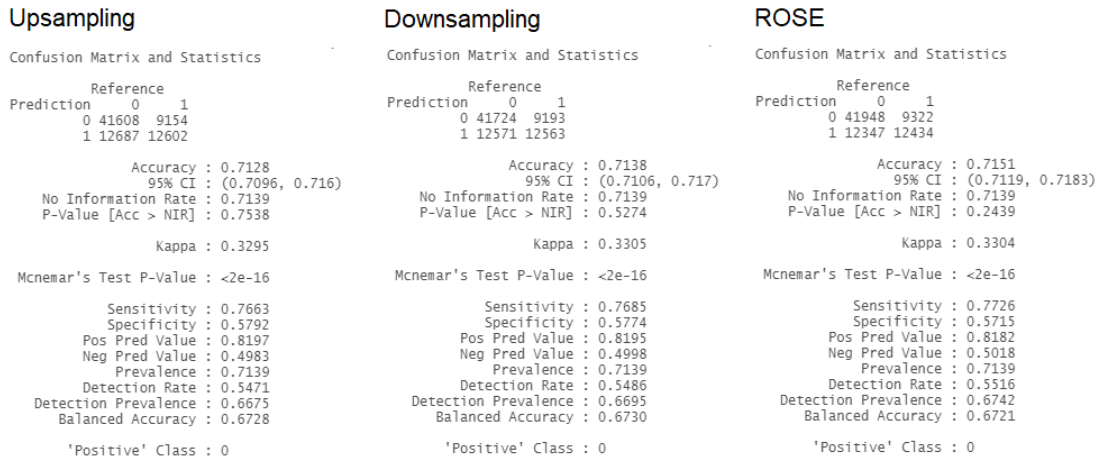


Figura 25: Modelos balanceados.

Se observa que los 3 modelos han mejorado el valor de la especificidad con respecto a los modelos anteriores en 30 puntos porcentuales. Como buscamos mejorar este valor lo máximo posible, se elegirá el modelo balanceado frente al no balanceado aunque se pierda poder de predicción (ha empeorado 4 puntos porcentuales).

¿Qué técnica de balanceo se elegirá? Como los resultados son muy similares, se va a elegir el que mayor especificidad devuelve. Esta es la técnica de upsampling (sobremuestreo).

En principio, ya se ha encontrado el modelo que se buscaba. Con un acierto de un 71.28 % y una especificidad de 57.92 %. Pero todavía se va a dar un paso más. Aunque se ha comentado que ante dos modelos con resultados similares,

por el principio de parsimonia, siempre se elige el más sencillo, como se tiene la intención de interpretar los parámetros del modelo para entender como se comportan las variables, se van a añadir algunas variables al modelo para poderlas estudiar y conocer como se comportan, aunque no aporten en exceso a la predicción.

Las variables que se añadirán al modelo además del **nivel crediticio** y **diferencia de prima** son: **perfil**, **año de renovación**, **paquete de garantías** y **club**.

8 Conclusiones y trabajo futuro

En este trabajo se han presentado los aspectos básicos para llevar a cabo un modelo de regresión logística binaria (formulación, evaluación, validación, etc.) Además se ha planteado como solucionar el problema de variables respuesta no balanceadas.

Para ponerlo en práctica, se ha seleccionado un conjunto de datos de una correduría especializada en motos, se han tenido que extraer los datos, hacer una limpieza de los mismos, estudiar las variables para ver cuáles interesaban para el modelo, se ha hecho una búsqueda del modelo óptimo empleando la técnica stepwise y diferentes técnicas de balanceo y por último se ha validado el modelo elegido, tanto con el conjunto de datos de prueba, como con un mes extraído de la base de datos que no habíamos incluido en el estudio.

Los resultados han determinado que un empeoramiento en el nivel crediticio del tomador repercute negativamente en la anulación. A diferencia de lo que sucede cuando aumentan los años del cliente en la aseguradora, o cuando tiene contratado un seguro más completo. En estos casos la anulación disminuye cuando aumentan estas variables.

Además se ha hecho uso del modelo, tanto para saber cuál es el efecto en la anulación tras aumentar la prima en X€, como para categorizar las probabilidades predichas de abandono en tres grupos (renueva, anula y duda), con el fin de realizar campañas de retención enfocadas en los clientes más indecisos.

Como líneas futuras se plantea incorporar la interacción entre variables en el modelo logístico para así poder obtener una mejor interpretación de como interactúan entre ellas.

Además se plantea el uso de otros modelos de clasificación como puede ser random forest o redes neuronales, ya que como se vio en algunos papers de la sección 3 estos métodos proporcionaban resultados muy interesantes.

Por último, en la sección 5 se explicó que se disponía de un total de 44 variables pero que en este trabajo no se iba a hacer uso de todas ellas.

Aunque se descartaran para este trabajo, eso no implica que no pudieran ser útil para dar respuesta a otras preguntas, como son:

- Estudiar la probabilidad de anulación dependiendo del regalo recibido. Y la probabilidad de anular si se ha recibido algún regalo, independientemente de cuál.
- De igual forma que en el punto anterior, pero con el uso de las coberturas.
- Modelo de regresión logística nominal para predecir si renovará, traspasará la póliza o anulará.
- Modelo de clasificación para realizar una categorización lógica de la variable provincia.

Referencias

[1] Alan Agresti. *Building and Applying Logistic Regression Models*. 2003, págs. 211-266.

[2] Hirotugu Akaike. «A new look at the statistical model identification». En: *IEEE transactions on automatic control* 19.6 (1974), págs. 716-723.

[3] Mohamed Boussiala. *Cook's Distance*. Oct. de 2020.

[4] George EP Box y Paul W Tidwell. «Transformation of the independent variables». En: *Technometrics* 4.4 (1962), págs. 531-550.

[5] Rocco Roberto Cerchiara, Matthew Edwards y Alessandra Gambini. «Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II». En: *18th international AFIR colloquium*. Citeseer. 2008.

[6] Nitesh V. Chawla y col. «SMOTE: synthetic minority over-sampling technique». En: *Journal of artificial intelligence research* 16 (2002), págs. 321-357.

[7] David Collett. *Modelling binary data*. CRC press, 2002.

[8] R Dennis Cook y Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman y Hall, 1982.

[9] David Roxbee Cox y E Joyce Snell. *Analysis of binary data*. Routledge, 2018.

[10] Michael Alin Efroymson. «Multiple regression analysis». En: *Mathematical methods for digital computers* (1960), págs. 191-203.

[11] Andreas Groll, Carsten Wasserfuhr y Leonid Zeldin. «Churn modeling of life insurance policies via statistical and machine learning methods–Analysis of important features». En: *arXiv preprint arXiv:2202.09182* (2022).

[12] Leo Guelman, Montserrat Guillén y Ana M Pérez-Marın. «Random forests for uplift modeling: an insurance customer retention case». En: *International conference on modeling and simulation in engineering, economics and management*. Springer. 2012, págs. 123-133.

[13] Ronald R Hocking. «A Biometrics invited paper. The analysis and selection of variables in linear regression». En: *Biometrics* (1976), págs. 1-49.

[14] David W Hosmer Jr, Stanley Lemeshow y Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[15] David G Kleinbaum y col. *Applied regression analysis and other multivariable methods*. Cengage Learning, 2013.

[16] Bart Larivière y Dirk Van den Poel. «Predicting customer retention and profitability by using random forests and regression forests techniques». En: *Expert systems with applications* 29.2 (2005), págs. 472-484.

[17] Daniel McFadden y col. «Conditional logit analysis of qualitative choice behavior». En: (1973).

[18] Nico JD Nagelkerke y col. «A note on a general definition of the coefficient of determination». En: *biometrika* 78.3 (1991), págs. 691-692.

[19] John Ashworth Nelder y Robert WM Wedderburn. «Generalized linear models». En: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), págs. 370-384.

[20] K Pearson. «Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242». En: *K Pearson* (1895).

[21] Robin L Plackett. «Karl Pearson and the chi-squared test». En: *International statistical review/revue internationale de statistique* (1983), págs. 59-72.

[22] Gideon Schwarz. «Estimating the dimension of a model». En: *The annals of statistics* (1978), págs. 461-464.

[23] Kate A Smith, Robert J Willis y Malcolm Brooks. «An analysis of customer retention and insurance claim patterns using data mining: A case study». En: *Journal of the operational research society* 51.5 (2000), págs. 532-541.

[24] Charles Spearman. «The proof and measurement of association between two things.» En: (1961).

[25] Tue Tjur. «Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination». En: *The American Statistician* 63.4 (2009), págs. 366-372.

9 Anexo

```
library(tidyr, warn.conflicts = FALSE)
library(caret, warn.conflicts = FALSE)
library(dplyr)     # Manejo de datos
library(ggplot2)   # Graficar
library(pROC)      # Curva ROC
library(DescTools) # Cálculo del coeficiente de determinación
library(ResourceSelection) # Test de Hosmer Lemeshow
library(coefplot)  # graficar coeficientes de un modelo
library(gmodels) # categorical variable
library(tinytex) # generar pdf con latex
library(fastDummies) # variables dummy
library(gridExtra) # Unir dos gráficos
library(RODBC) # Conexion base de datos
library(lubridate) # Fechas
library(ggpubr)    # Combinar gráficos
library(openxlsx)  # excel
library(Hmisc)
library(stargazer)
library(performance)
library(see)
library(readxl)
library(Hmisc)
library(car)
library(GGally)
library(lmtest)

################################################################################
############################## lectura de datos ################################
################################################################################

odbc_con <- odbcConnect("DWH01")
```

```
datos <- sqlQuery(odbc_con, "")


datos$PrimaVigorTotal <- round(datos$PrimaVigor + datos$PrimaVigorClub, 2)
datos$PrimaVigorTotalDiferencia <- round(datos$PrimaVigorDiferencia +
datos$PrimaVigorDiferenciaClub, 2)
datos <- datos[!is.na(datos$EdadAsegurado),]
datos <- datos[!is.na(datos$AniosCarnetTomador),]
datos <- datos[!is.na(datos$AniosMatriculacion),]



# variable independiente
datos$ANULA <- as.character(datos$ANULA)
datos$ANULA <- factor(datos$ANULA, levels = c("0", "1"),
labels = c("VIGOR", "ANULADA"))




datos$Club <- relevel(factor(datos$Club), ref = 'SIN CLUB')

datos$Mes <- relevel(factor(datos$Mes), ref = '1')

datos$Perfil <- relevel(factor(datos$Perfil), ref = '2 RECARGOS')

datos$CreditScoring <- relevel(factor(datos$CreditScoring,
levels = c("6", "5", "4", "3", "1 - 2", "A", "B", "C", "D", "F - E", "-")),
ref = '6')

#datos$CreditScoring <- factor(datos$CreditScoring, ordered = TRUE,
levels = c("6", "5", "4", "3", "1 - 2", "A", "B", "C", "D", "F - E", "-"))

datos$Accidentes <- relevel(factor(datos$Accidentes),
```

```
ref = 'SIN ACCIDENTES')

datos$PaqueteGarantias <- relevel(factor(datos$PaqueteGarantias),
ref = 'TERCEROS')

datos$Mediador <- factor(datos$Mediador)

datos$GrupoCartera <- factor(datos$GrupoCartera)

datos$Aceite <- factor(datos$Aceite)

datos$Kit_Care_Line <- factor(datos$Kit_Care_Line)

datos$TarjetaYamalube<- factor(datos$TarjetaYamalube)

datos$Buff <- factor(datos$Buff)

datos$Gafas <- factor(datos$Gafas)

datos$Chubasquero <- factor(datos$Chubasquero)

datos$Cheque_regalo_Kawa <- factor(datos$Cheque_regalo_Kawa)

datos$CambioAceiteYamalube <- factor(datos$CambioAceiteYamalube)

datos$SubsidioPorHospitalizacion <-
factor(datos$SubsidioPorHospitalizacion)

datos$GestionDeMultas <- factor(datos$GestionDeMultas)

datos$AsistentePersonalDependencia <-
factor(datos$AsistentePersonalDependencia)

datos$RentACar <- factor(datos$RentACar)
```

```
datos$ITV <- factor(datos$ITV)

datos$SustitucionNeumatico <- factor(datos$SustitucionNeumatico)

datos$AveriaMecanica <- factor(datos$AveriaMecanica)

datos$VehiculoSustAccidente <- factor(datos$VehiculoSustAccidente)

datos$VehiculoSustAveria <- factor(datos$VehiculoSustAveria)

datos$Llaves <- factor(datos$Llaves)

datos$EquipamientoMotorista <- factor(datos$EquipamientoMotorista)

datos$SeguroDesempleo <- factor(datos$SeguroDesempleo)

datos$IncapacidadTemporal <- factor(datos$IncapacidadTemporal)

datos$ReparacionNeumatico <- factor(datos$ReparacionNeumatico)

datos$AsistenciaJuridicaTelefonica <-
factor(datos$AsistenciaJuridicaTelefonica)

datos$AsistenciaMecanicaTelefonica <-
factor(datos$AsistenciaMecanicaTelefonica)

datos$TV <- factor(datos$TV)

datos$CC <- factor(datos$CC)


datos$marca <- as.character(datos$marca)
lista_marcas <- datos %>% group_by(marca) %>%
```

```
summarise(cuenta = n()) %>% arrange(desc(cuenta)) %>%
filter(cuenta >= 1000) %>% pull(marca)
datos$marca <- as.factor(case_when(
  datos$marca %in% lista_marcas ~ datos$marca,
  TRUE ~ "RESTO"
))


datos$AnyoRenovacionComparativa <-
as.integer(datos$AnyoRenovacionComparativa)

datos$Division <- factor(datos$Division)

datos$Av <- factor(datos$Av)

datos$Def <- factor(datos$Def)

datos$Provincia <- sprintf("%02d", datos$Provincia)
datos$Provincia <- ifelse(datos$Provincia == "00", "NA", datos$Provincia)

datos$fr_edad <- as.factor(case_when(
  datos$EdadAsegurado <= 17 ~ "1. Menores de 18",
  datos$EdadAsegurado <= 20 ~ "2. Menores de 21",
  datos$EdadAsegurado <= 24 ~ "3. Menores de 25",
  datos$EdadAsegurado <= 29 ~ "4. Menores de 30",
  datos$EdadAsegurado <= 37 ~ "5. Menores de 38",
  datos$EdadAsegurado <= 55 ~ "6. Menores de 56",
  datos$EdadAsegurado >= 56 ~ "7. 56 o mas"))

datos$fr_carnet <- as.factor(case_when(
  datos$AniosCarnetTomador <= 1 ~ "1. 0 y 1",
  datos$AniosCarnetTomador <= 5 ~ "2. Menores de 6",
  datos$AniosCarnetTomador <= 40 ~ "3. Menores de 41",
  datos$AniosCarnetTomador <= 100 ~ "4. 41 o mas"))
```

```
datos$fr_matricula <- as.factor(case_when(
  datos$AniosMatriculacion <= 1 ~ "1. 0 y 1",
  datos$AniosMatriculacion <= 14 ~ "2. Menores de 15",
  datos$AniosMatriculacion <= 19 ~ "3. Menores de 20",
  datos$AniosMatriculacion <= 100 ~ "4. 20 o mas"))

datos$fr_provincia <- as.factor(case_when(
  datos$Provincia %in% c('08', '27', '07', '51') ~ "1. Zona",
  datos$Provincia %in% c('17', '15', '29', '28', '31', '40', '33',
  '43', '20', '48', '46', '41') ~ "2. Zona", # Menores de 31% anulacion
  datos$Provincia %in% c('01', '03', '04', '11', '12', '18', '19',
  '21', '22', '24', '25', '26', '32', '35', '36', '37', '38', '39',
  '42', '44', '47', '49', '50', '52') ~ "3. Zona", # Menores de 36% anulacion
  datos$Provincia %in% c('09', '45', '13', '16', '05', '34', '02',
  '30', '14', '10', '06', '23') ~ "4. Zona")) # Menores de 51% anulacion


################################################################################
############################## descripción variables ###########################
################################################################################

describe <- function (df, varib, ajuste) {

  columna <- as.symbol(varib)

  resumen <- df %>%
    group_by(!!columna) %>%
    summarise(incidencia = n(),
    conversion = round(sum(ifelse(ANULA=='ANULADA', 1, 0))*100/n(),1))
```

```
  g2 <- ggplot(resumen, aes(x = factor(!!columna))) +
    geom_col(aes(y = incidencia), fill = "#d6dce9", alpha = 0.5) +
    geom_line(aes(y = conversion * ajuste), group = 1, color = "#58517a") +
    #geom_text(aes(y = conversion * ajuste, label = conversion),
    color = "#58517a",
    size = 2, vjust = 1.8, fontface = 'bold') +
    scale_y_continuous(sec.axis = sec_axis(~./ajuste), name = "") +
    theme(text = element_text(size = 14),
          axis.text.x = element_text(angle = 90, hjust = 1, size = 1),
          plot.title = element_text(size = 2, face = "bold")) +
    theme_bw() +
    labs(y = NULL, x = NULL, title = paste('Distribución del % de anulación
    en la variable:', varib))
  g2
}



describe(datos, 'EdadAsegurado', 100)
describe(datos, 'AniosCarnetTomador', 100)
describe(datos, 'AniosMatriculacion', 700)


#### Regalos

p1 <- ggplot(datos, aes(Aceite)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = 'Aceite') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))
```

```
p2 <- ggplot(datos, aes(Kit_Care_Line)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = 'Kit Care Line') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
             vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))

p3 <- ggplot(datos, aes(TarjetaYamalube)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = 'Tarjeta Yamalube') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
             vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))

p4 <- ggplot(datos, aes(Gafas)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = 'Gafas') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
             vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))

p5 <- ggplot(datos, aes(Chubasquero)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
```

```
  labs(y = "Incidencia", x = NULL, title = 'Chubasquero') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))

p6 <- ggplot(datos %>% mutate(Cheque_regalo_Kawa =
ifelse(Cheque_regalo_Kawa == "NO APLICA", "NO",
ifelse(Cheque_regalo_Kawa == "NO", "NO", "SI"))), aes(Cheque_regalo_Kawa)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = 'Cheque Regalo Kawasaki') +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.5) +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 9, face = "bold"))

ggarrange(p1, p2, p3, p4, p5, p6,
          ncol = 3, nrow = 2)


#### Uso Garantias

p1 <- ggplot(datos, aes(CambioAceiteYamalube)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Cambio de aceite') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p2 <- ggplot(datos, aes(SubsidioPorHospitalizacion)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
```

```
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Subsidio por hospitalización') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p3 <- ggplot(datos, aes(GestionDeMultas)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Gestión de multas') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p4 <- ggplot(datos, aes(AsistentePersonalDependencia)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Asistente personal por dependencia') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p5 <- ggplot(datos, aes(RentACar)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Rent A Car') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p6 <- ggplot(datos, aes(ITV)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'ITV') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p7 <- ggplot(datos, aes(SustitucionNeumatico)) +
```

```
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Sustitución de neumáticos') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p8 <- ggplot(datos, aes(AveriaMecanica)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Avería mecánica') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p9 <- ggplot(datos, aes(VehiculoSustAccidente)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Vehiculo de sustitución por accidente') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p10 <- ggplot(datos, aes(VehiculoSustAveria)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Vehiculo de sustitución por avería') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p11 <- ggplot(datos, aes(Llaves)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Llaves') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))
```

```
p12 <- ggplot(datos, aes(EquipamientoMotorista)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Equipamiento al motorista') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p13 <- ggplot(datos, aes(SeguroDesempleo)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Seguro por desempleo') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p14 <- ggplot(datos, aes(IncapacidadTemporal)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Incapacidad temporal') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p15 <- ggplot(datos, aes(ReparacionNeumatico)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Reparación de neumáticos') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

p16 <- ggplot(datos, aes(AsistenciaJuridicaTelefonica)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Asistencia jurídica telefónica') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))
```

```
p17 <- ggplot(datos, aes(AsistenciaMecanicaTelefonica)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = NULL, x = NULL, title = 'Asistencia mecánica telefónica') +
  ylim(0, 250000) +
  theme(plot.title = element_text(size = 5, face = "bold"))

ggarrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, p13,
p14, p15, p16, p17, ncol = 5, nrow = 4)


#### Provincia

p1 <- ggplot(datos, aes(fr_provincia)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
p2 <- datos %>% group_by(fr_provincia, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(fr_provincia) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = fr_provincia, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 3) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
ggarrange(p1, p2,
          ncol = 2, nrow = 1)


#### Resto


p1 <- datos %>% group_by(Mes, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Mes) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Mes, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "Mes") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(AnyoRenovacionComparativa, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(AnyoRenovacionComparativa) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = AnyoRenovacionComparativa, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "AnyoRenovacionComparativa") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
p3 <- datos %>% group_by(TV, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(TV) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = TV, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "TV") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p4 <- datos %>% group_by(CreditScoring, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(CreditScoring) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = CreditScoring, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "CreditScoring") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2, p3, p4,
          ncol = 2, nrow = 2)

p1 <- datos %>% group_by(Mediador, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Mediador) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Mediador, y = Percentage, fill = ANULA)) +
```

```
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "Mediador") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(Club, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Club) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Club, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "Club") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p3 <- datos %>% group_by(CC, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(CC) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = CC, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "CC") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
p4 <- datos %>% group_by(PaqueteGarantias, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(PaqueteGarantias) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = PaqueteGarantias, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "PaqueteGarantias") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p5 <- datos %>% group_by(Perfil, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Perfil) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Perfil, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "Perfil") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p6 <- datos %>% group_by(GrupoCartera, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(GrupoCartera) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = GrupoCartera, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
```

```
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 2) +
  labs(x = NULL, title = "GrupoCartera") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2, p3, p4, p5, p6,
          ncol = 3, nrow = 2)



### ANULA


ggplot(datos, aes(ANULA)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL, title = "Estado póliza") +
  ylim(0, 170000) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.6)


### Club


p1 <- ggplot(datos, aes(Club)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
```

```
            vjust = -0.2)

p2 <- datos %>% group_by(Club, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Club) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Club, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

datos %>% group_by(Mediador, Anualidad, Club, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(Mediador, Anualidad, Club) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Club, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  facet_grid(Mediador~Anualidad)


### TV
```

```
p1 <- ggplot(datos, aes(TV)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(TV, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(TV) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = TV, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 3) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)



### CC


p1 <- ggplot(datos, aes(CC)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
```

```
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(CC, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(CC) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = CC, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 3) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### marca


p1 <- ggplot(datos, aes(marca)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", angle = 90,
  vjust = 0.5, hjust = -0.3, fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
```

```
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(marca, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(marca) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = marca, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)), angle = 90,
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### Division


p1 <- ggplot(datos, aes(Division)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", angle = 90,
  vjust = 0.5, hjust = -0.3, fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
  hjust=1))
```

```
p2 <- datos %>% group_by(Division, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(Division) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Division, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)), angle = 90,
  size = 2.8, position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5,
        hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### Accidentes


p1 <- ggplot(datos, aes(Accidentes)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
  hjust=1))

p2 <- datos %>% group_by(Accidentes, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
```

```
group_by(Accidentes) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Accidentes, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5,
        hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### Av


p1 <- ggplot(datos, aes(Av)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.2)

p2 <- datos %>% group_by(Av, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Av) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Av, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
```

```
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### Def


p1 <- ggplot(datos, aes(Def)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2)

p2 <- datos %>% group_by(Def, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Def) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Def, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
```

```
          ncol = 2, nrow = 1)



### PaqueteGarantias


p1 <- ggplot(datos, aes(PaqueteGarantias)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
  hjust=1))

p2 <- datos %>% group_by(PaqueteGarantias, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(PaqueteGarantias) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = PaqueteGarantias, y = Percentage,
  fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), size = 3.5) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90,
        vjust = 0.5, hjust=1))

ggarrange(p1, p2,
```

```
          ncol = 2, nrow = 1)


### Perfil


p1 <- ggplot(datos, aes(Perfil)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(Perfil, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Perfil) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Perfil, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

datos %>% group_by(Division, Perfil, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(Division, Perfil) %>%
```

```
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Perfil, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  facet_wrap(~Division)


### CreditScoring


p1 <- ggplot(datos, aes(CreditScoring)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 2.5, aes(label = ..count..), position = position_dodge(1),
              vjust = -0.2)

p2 <- datos %>% group_by(CreditScoring, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(CreditScoring) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = CreditScoring, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)), angle = 90,
  position = position_stack(.5)) +
```

```
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

datos %>% group_by(Anualidad, CreditScoring, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(Anualidad, CreditScoring) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = CreditScoring, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  facet_wrap(~Anualidad)


### GrupoCartera


p1 <- ggplot(datos, aes(GrupoCartera)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 4, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.2)

p2 <- datos %>% group_by(GrupoCartera, ANULA) %>%
```

```
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(GrupoCartera) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = GrupoCartera, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### Mes


p1 <- ggplot(datos, aes(Mes)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", angle = 90, vjust = 0.4,
  hjust = -0.5, colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
              vjust = -0.2)

p2 <- datos %>% group_by(Mes, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(Mes) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = Mes, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
```

```
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), angle = 90, size = 3) +
  labs(x = NULL) +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### AnyoRenovacionComparativa


p1 <- ggplot(datos, aes(AnyoRenovacionComparativa)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", angle = 90, vjust = 0.4,
  hjust = -0.5, colour = "black", fontface = "bold", size = 3,
  aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2)

p2 <- datos %>% group_by(AnyoRenovacionComparativa, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(AnyoRenovacionComparativa) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = AnyoRenovacionComparativa, y = Percentage,
  fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5), angle = 90, size = 3) +
  labs(x = NULL) +
```

```
  theme(legend.position = "top")


ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### AniosMatriculacion


p1 <- ggplot(datos, aes(AniosMatriculacion)) +
  geom_histogram(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw()

p2 <- ggplot(datos, aes(AniosMatriculacion, fill = ANULA)) +
  geom_histogram(colour = "black", lwd = 0.25, linetype = 1, alpha = 0.8,
  position = "identity") +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

ggplot(datos) +
  geom_boxplot(aes(x = AniosMatriculacion, y = factor(ANULA),
  colour = factor(ANULA))) +
  theme_bw() + theme(legend.position = "none") +
  labs(title = "Boxplot: Años desde la matriculacion vs Estado Poliza") +
  scale_color_manual(values = c("#d6dce9", "#58517a")) + labs(y = NULL)


### EdadAsegurado
```

```
p1 <- ggplot(datos, aes(EdadAsegurado)) +
  geom_histogram(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw()

p2 <- ggplot(datos, aes(EdadAsegurado, fill = ANULA)) +
  geom_histogram(colour = "black", lwd = 0.25, linetype = 1,
  alpha = 0.8, position = "identity") +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

ggplot(datos) +
  geom_boxplot(aes(x = EdadAsegurado, y = factor(ANULA),
  colour = factor(ANULA))) +
  theme_bw() + theme(legend.position = "none") +
  labs(title = "Boxplot: Años desde la matriculacion vs Estado Poliza") +
  scale_color_manual(values = c("#d6dce9", "#58517a")) + labs(y = NULL)


### AniosCarnetTomador


p1 <- ggplot(datos, aes(AniosCarnetTomador)) +
  geom_histogram(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw()

p2 <- ggplot(datos, aes(AniosCarnetTomador, fill = ANULA)) +
  geom_histogram(colour = "black", lwd = 0.25, linetype = 1,
  alpha = 0.8, position = "identity") +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
```

```
  theme_bw() +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

ggplot(datos) +
  geom_boxplot(aes(x = AniosCarnetTomador, y = factor(ANULA),
  colour = factor(ANULA))) +
  theme_bw() + theme(legend.position = "none") +
  labs(title = "Boxplot: Años desde la matriculacion vs Estado Poliza") +
  scale_color_manual(values = c("#d6dce9", "#58517a")) + labs(y = NULL)



### fr_edad


p1 <- ggplot(datos, aes(fr_edad)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
            vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(fr_edad, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(fr_edad) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = fr_edad, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
```

```
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### fr_carnet


p1 <- ggplot(datos, aes(fr_carnet)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(fr_carnet, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>% group_by(fr_carnet) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = fr_carnet, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
ggarrange(p1, p2,
          ncol = 2, nrow = 1)


### fr_matricula


p1 <- ggplot(datos, aes(fr_matricula)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw() +
  labs(y = "Incidencia", x = NULL) +
  stat_count(geom = "text", colour = "black", fontface = "bold",
  size = 3, aes(label = ..count..), position = position_dodge(1),
             vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

p2 <- datos %>% group_by(fr_matricula, ANULA) %>%
summarise(Percentage = n(), .groups = 'drop') %>%
group_by(fr_matricula) %>%
  mutate(Percentage = Percentage/sum(Percentage)*100) %>%
  ggplot(aes(x = fr_matricula, y = Percentage, fill = ANULA)) +
  geom_col(color = "black", alpha = 0.9) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  geom_text(aes(label = round(Percentage, 2)),
  position = position_stack(.5)) +
  labs(x = NULL) +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

ggarrange(p1, p2,
          ncol = 2, nrow = 1)
```

```
### PrimaVigorTotal


p1 <- ggplot(datos, aes(PrimaVigorTotal)) +
  geom_histogram(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw()

p2 <- ggplot(datos, aes(PrimaVigorTotal, fill = ANULA))+
  geom_density(alpha = 0.7) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

ggplot(datos) +
  geom_boxplot(aes(x = PrimaVigorTotal, y = factor(ANULA),
  colour = factor(ANULA))) +
  theme_bw() + theme(legend.position = "none") +
  labs(title = "Boxplot: Diferencia de prima total vs Estado Poliza") +
  scale_color_manual(values = c("#d6dce9", "#58517a")) + labs(y = NULL)



### PrimaVigorTotalDiferencia


p1 <- ggplot(datos, aes(PrimaVigorTotalDiferencia)) +
  geom_histogram(fill = "#58517a", colour = "black", alpha = 0.9) +
  theme_bw()
```

```
p2 <- ggplot(datos, aes(PrimaVigorTotalDiferencia, fill = ANULA))+
  geom_density(alpha = 0.7) +
  scale_fill_manual(values = c("#d6dce9", "#58517a")) +
  theme_bw() +
  theme(legend.position = "top")

ggarrange(p1, p2,
          ncol = 2, nrow = 1)

ggplot(datos) +
  geom_boxplot(aes(x = PrimaVigorTotalDiferencia,
  y = factor(ANULA), colour = factor(ANULA))) +
  theme_bw() + theme(legend.position = "none") +
  labs(title = "Boxplot: Diferencia de prima total vs Estado Poliza") +
  scale_color_manual(values = c("#d6dce9", "#58517a")) + labs(y = NULL)


################################################################################
################################ correlaciones ################################
################################################################################

# variables seleccionadas para el modelo
data <- datos %>% select(Mes, Mediador, marca, Division,
AnyoRenovacionComparativa, Club, TV, CC, Accidentes, Av, Def,
PaqueteGarantias, Perfil, CreditScoring, PrimaVigorTotal,
PrimaVigorTotalDiferencia, fr_edad, fr_carnet, fr_matricula,
fr_provincia, ANULA)



# Select categorical column
factor <- data.frame(select_if(data, is.factor))

# Convert data to numeric
```

```
corr <- data.frame(lapply(factor, as.integer))

# Plot the graph
ggcorr(corr,
       method = c("pairwise", "pearson"),
       nbreaks = 9,
       low = "#572e3b", mid = "#e7eaf2", high = "#3b572e", # color
       label = TRUE, label_size = 2.5, label_color = "black", # label
       hjust = 0.9, size = 2.2, color = "grey20") # variable label




# Select integer column
integer <- data.frame(data$PrimaVigorTotal,
data$PrimaVigorTotalDiferencia,
data$AnyoRenovacionComparativa, data$ANULA)

# Convert data to numeric
corr <- data.frame(lapply(integer, as.integer))

# Plot the graph
ggcorr(corr,
       method = c("pairwise", "spearman"),
       nbreaks = 5,
       low = "#572e3b", mid = "#e7eaf2", high = "#3b572e", # color
       label = TRUE, label_size = 2.5, label_color = "black", # label
       hjust = 0.7, size = 3, color = "grey20") # variable label


################################################################################
################################### modelado ###################################
################################################################################
```

```
df <- datos %>% select(Mes, Mediador, marca, Division,
AnyoRenovacionComparativa, Club, TV, CC, Accidentes, Av, Def,
PaqueteGarantias, Perfil, CreditScoring, PrimaVigorTotal,
PrimaVigorTotalDiferencia, fr_edad,fr_carnet, fr_matricula,
fr_provincia, ANULA)


set.seed(3456)
trainIndex <- createDataPartition(df$ANULA, p = 0.66,
                                  list = FALSE,
                                  times = 1)
data_train <- df[trainIndex,]
data_test  <- df[-trainIndex,]

table(data_train$ANULA)




# modelos simples sin balancear
logit_1 <- glm(ANULA ~ CreditScoring, data = data_train,
family = 'binomial')
predict_1 <- predict(logit_1, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_1>0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))

logit_2 <- glm(ANULA ~ PrimaVigorTotalDiferencia, data = data_train,
family = 'binomial')
predict_2 <- predict(logit_2, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_2>0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))

logit_3 <- glm(ANULA ~ Perfil, data = data_train, family = 'binomial')
```

```
summary(logit_3)
histogram(predict(logit_3, data_test, type = 'response'))
predict_3 <- predict(logit_3, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_3>0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))

# modelos con dos variables sin balancear
logit_4 <- glm(ANULA ~ Perfil + CreditScoring, data = data_train,
family = 'binomial')
predict_4 <- predict(logit_4, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_4>0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))


logit_5 <- glm(ANULA ~ CreditScoring + PrimaVigorTotalDiferencia,
data = data_train, family = 'binomial')
predict_5 <- predict(logit_5, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_5>0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))




# StepWise sin balancear
logit_0 <- glm(ANULA ~ 1, data = data_train, family = "binomial")

logit_full <- glm(ANULA ~ Mes + Mediador + marca + Division +
AnyoRenovacionComparativa + Club + TV + CC + Accidentes + Av +
Def + PaqueteGarantias + Perfil + CreditScoring + PrimaVigorTotal +
PrimaVigorTotalDiferencia + fr_edad + fr_carnet +
fr_matricula + fr_provincia,
data = data_train, family = "binomial")
```

```
predict_step <- predict(logit_step, data_test, type = 'response')
confusionMatrix(data = as.factor(as.integer(predict_step > 0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))
```

```
# 2 variables balanceado: sobremuestreo

set.seed(3456)

# upSample para realizar la técnica de muestreo superior.
trainup <- upSample(x = data_train[, -ncol(data_train)],
                    y = data_train$ANULA,
                    yname = "ANULA")

table(trainup$ANULA)

modelup <- glm(ANULA ~ CreditScoring + PrimaVigorTotalDiferencia,
data = trainup, family = "binomial")

summary(modelup)

predict_trainup <- predict(modelup, data_test, type = 'response')

# histograma probabilidades
histogram(predict(modelup, data_test, type = 'response'))

confusionMatrix(data = as.factor(as.integer(predict_trainup > 0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))
```

```
# 2 variables balanceado: inframuestreo

set.seed(3456)

# upSample para realizar la técnica de muestreo superior.
traindown <- downSample(x = data_train[, -ncol(data_train)],
                        y = data_train$ANULA,
                        yname = "ANULA")

table(traindown$ANULA)

modeldown <- glm(ANULA ~ CreditScoring + PrimaVigorTotalDiferencia,
data = traindown,
family = "binomial")

summary(modeldown)

predict_traindown <- predict(modeldown, data_test, type = 'response')

# histograma probabilidades
histogram(predict(modeldown, data_test, type = 'response'))

confusionMatrix(data = as.factor(as.integer(predict_traindown > 0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))




# 2 variables balanceado: ROSE

library(ROSE)
```

```
set.seed(3456)

# upSample para realizar la técnica de muestreo superior.
trainROSE <- ROSE(ANULA ~ ., data = data_train)$data

table(trainROSE$ANULA)

modelROSE <- glm(ANULA ~ CreditScoring + PrimaVigorTotalDiferencia,
data = trainROSE, family = "binomial")

summary(modelROSE)

predict_ROSE <- predict(modelROSE, data_test, type = 'response')

# histograma probabilidades
histogram(predict(modelROSE, data_test, type = 'response'))

confusionMatrix(data = as.factor(as.integer(predict_ROSE > 0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))




# 6 variables BALANCEADO: sobremuestreo

set.seed(3456)

# upSample para realizar la técnica de muestreo superior.
trainup <- upSample(x = data_train[, -ncol(data_train)],
                    y = data_train$ANULA,
                    yname = "ANULA")

table(trainup$ANULA)
```

```
model_final <- glm(ANULA ~ CreditScoring + PrimaVigorTotalDiferencia +
Perfil + AnyoRenovacionComparativa + PaqueteGarantias + Club,
data = trainup,
family = "binomial")

summary(model_final)

predict_final <- predict(model_final, data_test, type = 'response')

# histograma probabilidades
histogram(predict(model_final, data_test, type = 'response'),
layout = c(2, 3))

confusionMatrix(data = as.factor(as.integer(predict_final > 0.5)),
    reference = as.factor(ifelse(data_test$ANULA == "VIGOR", 0, 1)))




##############################################################################
############################## supuestos del modelo ##########################
##############################################################################

# observaciones independientes
plot(residuals(model_final))

#Multicolinealidad
vif(model_final)

#valores atipicos extremos
plot(model_final, 4)
plot(model_final, 5)
```

```
#########################################################################
############################ bondad de ajuste ###########################
#########################################################################

# test hosmer y lemeshow

hosmerlem.test <- function(y, yhat, g = 10, group = F){

  colnum <- ncol(y)

  if(group == F){

    cutyhat1 = cut(yhat, breaks = unique(quantile(yhat,
    probs = seq(0, 1, 1/g))),
    include.lowest = TRUE)
    obs = xtabs(cbind(1 - y[, colnum], y[, colnum]) ~ cutyhat1)
    expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat1)

  }

  else{

    y2 <- c(rep(seq(0, 0, length = nrow(y)), y[, colnum - 1]),
    rep(seq(1, 1, length = nrow(y)), y[, colnum]))
    yhat2 <- c(rep(yhat, y[, colnum - 1]), rep(yhat, y[, colnum]))
    cutyhat1 = cut(yhat2, breaks = unique(quantile(yhat2,
    probs = seq(0, 1, 1/g))), include.lowest = TRUE)
    obs = xtabs(cbind(1 - y2, y2) ~ cutyhat1)
    expect = xtabs(cbind(1 - yhat2, yhat2) ~ cutyhat1)

  }
```

```
  chisq.C = sum((obs - expect)^2/expect)
  P.C = 1 - pchisq(chisq.C, g - 2)
  res <- data.frame(c(chisq.C, P.C))
  colnames(res) <- c("Hosmer-Lemeshow Test")
  rownames(res) <- c("X-squared","p.value")
  return(res)
}


hosmerlem.test(trainup, fitted.values(model_final), g=10, group=F)

# likelihood

logit_null <- glm(ANULA ~ 1, data = trainup, family = "binomial")
lrtest(model_final, logit_null)


# Deviance (ratio-likelihood)
pchisq(38329, 31, lower.tail = FALSE)


### R-square
with(summary(model_final), 1 - deviance/null.deviance)



# La curva ROC

CurvaROC <- roc(data_test$ANULA, predict_final)
CurvaROC

#calculate auc
auc <- round(auc(data_test$ANULA, predict_final), 4)
```

```
#create ROC plot
ggroc(CurvaROC, colour = '#58517a', size = 2) + theme_bw() +
  ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))


#########################################################################
############################## coeficientes, ODDS ######################
#########################################################################


formattable(cbind(OR = coef(model_final),
            `cociente ventajas` = exp(coef(model_final)),
            exp(confint.default(model_final))))



trainup %>%
    select(CreditScoring, PrimaVigorTotal, Perfil,
    AnyoRenovacionComparativa,
    PaqueteGarantias, Club) %>%
  mutate(prob_exito=fitted.values(model_final),
 cociente_ventajas_1=
 fitted.values(model_final)/(1-fitted.values(model_final)),
 cociente_ventajas_0=
 (1-fitted.values(model_final))/fitted.values(model_final)) %>%
 unique() %>% filter(Trabajo=="Hosteleria" & Contactos=="Trabajo")




#########################################################################
################################### abril ###############################
#########################################################################
```

```
abril <- sqlQuery(odbc_con, "")


abril$PrimaVigorTotal <- round(abril$PrimaVigor + abril$PrimaVigorClub, 2)
abril$PrimaVigorTotalDiferencia <-
    round(abril$PrimaVigorDiferencia + abril$PrimaVigorDiferenciaClub, 2)
abril <- abril[!is.na(abril$EdadAsegurado),]
abril <- abril[!is.na(abril$AniosCarnetTomador),]
abril <- abril[!is.na(abril$AniosMatriculacion),]

abril$ANULA <- as.character(abril$ANULA)
abril$ANULA <- factor(abril$ANULA, levels = c("0", "1"),
                      labels = c("VIGOR", "ANULADA"))


# variables dependientes, ordinales
abril$Club <- relevel(factor(abril$Club), ref = 'SIN CLUB')

abril$Mes <- factor(abril$Mes)

abril$Perfil <- relevel(factor(abril$Perfil), ref = '2 RECARGOS')

abril <- abril %>%
    mutate(CreditScoring = ifelse(CreditScoring == "44593", "1 - 2",
                                  CreditScoring))
abril$CreditScoring <- relevel(factor(abril$CreditScoring,
    levels = c("6", "5", "4", "3", "1 - 2", "A",
    "B", "C", "D", "F - E", "-")),
    ref = '6')

abril$Accidentes <- relevel(factor(abril$Accidentes),
ref = 'SIN ACCIDENTES')
```

```
# variables dependientes, nominales
abril$Mediador <- factor(abril$Mediador)

abril$Anualidad <- factor(abril$Anualidad)

abril$GrupoCartera <- factor(abril$GrupoCartera)

abril$ProduccionCartera <- factor(abril$ProduccionCartera)

abril$TV <- factor(abril$TV)

abril$CC <- ifelse(abril$CC == "125", "= 125", abril$CC)
abril$CC <- factor(abril$CC)


abril$marca <- as.character(abril$marca)
abril$marca <- as.factor(case_when(
  abril$marca %in% lista_marcas ~ abril$marca,
  TRUE ~ "RESTO"
))


abril$AnyoRenovacionComparativa <-
    as.integer(abril$AnyoRenovacionComparativa)

abril$Division <- factor(abril$Division)

abril$Av <- factor(abril$Av)

abril$Def <- factor(abril$Def)

abril$PaqueteGarantias <- relevel(factor(abril$PaqueteGarantias),
```

```
                                    ref = 'TERCEROS')

abril$fr_edad <- as.factor(case_when(
  abril$EdadAsegurado <= 17 ~ "1. Menores de 18",
  abril$EdadAsegurado <= 20 ~ "2. Menores de 21",
  abril$EdadAsegurado <= 24 ~ "3. Menores de 25",
  abril$EdadAsegurado <= 29 ~ "4. Menores de 30",
  abril$EdadAsegurado <= 37 ~ "5. Menores de 38",
  abril$EdadAsegurado <= 55 ~ "6. Menores de 56",
  abril$EdadAsegurado >= 56 ~ "7. 56 o mas"))

abril$fr_carnet <- as.factor(case_when(
  abril$AniosCarnetTomador <= 1 ~ "1. 0 y 1",
  abril$AniosCarnetTomador <= 5 ~ "2. Menores de 6",
  abril$AniosCarnetTomador <= 40 ~ "3. Menores de 41",
  abril$AniosCarnetTomador <= 100 ~ "4. 41 o mas"))

abril$fr_matricula <- as.factor(case_when(
  abril$AniosMatriculacion <= 1 ~ "1. 0 y 1",
  abril$AniosMatriculacion <= 14 ~ "2. Menores de 15",
  abril$AniosMatriculacion <= 19 ~ "3. Menores de 20",
  abril$AniosMatriculacion <= 100 ~ "4. 20 o mas"))

abril$fr_provincia <- as.factor(case_when(
  abril$Provincia %in% c('08', '27', '07', '51') ~ "1. Zona",
  abril$Provincia %in% c('17', '15', '29', '28', '31', '40', '33',
        '43', '20', '48',
        '46', '41') ~ "2. Zona", # Menores de 31% anulacion
  abril$Provincia %in% c('36', '01', '25', '42', '44', '52', '12',
        '19', '49', '38', '22', '32', '26', '50', '35', '47', '37',
        '18', '24', '21', '03', '11', '04',
        '39') ~ "3. Zona", # Menores de 36% anulacion
  abril$Provincia %in% c('09', '45', '13', '16', '05', '34', '02',
  '30', '14', '10', '06', '23') ~ "4. Zona")) # Menores de 51% anulacion
```

```
predict_abril <- predict(model_final, abril, type = 'response')

# histograma probabilidades
histogram(predict(model_final, abril, type = 'response'))

confusionMatrix(data = as.factor(as.integer(predict_abril>0.5)),
                reference = as.factor(ifelse(abril$ANULA == "VIGOR", 0, 1)))



a <- data.frame(predict_abril, abril$ANULA)
a$categoria <- ifelse(predict_abril < 0.3, "Renueva",
                ifelse(predict_abril>0.7, "Anula", "Duda"))

ggplot(a, aes(x = categoria)) +
  geom_bar(fill = "#58517a", colour = "black", alpha = 0.9)+
  theme_bw() +
  labs(y = "Incidencia", x = NULL,
    title = "Categorización propuesta en base a la predicción") +
  stat_count(geom = "text", colour = "black",
             fontface = "bold", size = 4, aes(label = ..count..),
             position = position_dodge(1),
             vjust = -0.2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ylim(0, 5000)

table(a$abril.ANULA, a$categoria)
```