UNIVERSIDAD ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

MODELO PREDICTIVO DE RETENCIÓN DE CLIENTES DE CARTERA PARA EL SEGURO DE AUTOMÓVILES

MAESTRÍA EN CIENCIAS ACTUARIALES

AUTOR: ANGELA PAOLA TIBOCHE RUBIANO

2023-II

MODELO PREDICTIVO DE RETENCIÓN DE CLIENTES DE CARTERA PARA EL SEGURO DE AUTOMÓVILES

Autor: Angela Paola Tiboche Rubiano

Tutores: Roberto Perez – Nicholas Metaxas

Universidad Escuela Colombiana de Ingeniería Julio Garavito

Maestría en Ciencias Actuariales, 2023 – II

Yo Angela Tiboche declaro que el contenido de este documento es de propiedad intelectual mía, y no posee plagio de algún tipo, además las referencias de este han sido citadas al final del texto.

RESUMEN

Las compañías aseguradoras cada día le dan más importancia a la recolección de datos del negocio, el hecho de almacenar cada una de las variables relacionadas al riesgo asegurado nos permite mejorar los procesos y el desarrollo de proyectos de optimización, los cuales aportan valor a cada una de las operaciones del negocio, en especial a aquellos ramos que modelen su tarifa a partir de modelos multivariables (Modelos Lineales Generalizados), pues esto le permite tener una tarifa mejor perfilada para cada uno de los clientes y tener tarifas más competitivas en el mercado.

Los esfuerzos del área comercial para poder ingresar nueva producción a la compañía son muy grandes, e incurren en riesgo de llegar a tener procesos de anti-selección a clientes con el único fin de generar ingresos en primas a la compañía. Sin embargo, la cartera de clientes histórica es uno de los mayores activos con los que cuenta cualquier compañía de seguros, lo cual debería de generar mayor interés en la retención de clientes, pues esto mejora los resultados técnicos de la línea de negocio correspondiente. El hecho de que un cliente cancele un contrato de seguro implica la oportunidad de generar beneficios futuros, aún más si en sus vigencias anteriores no incurrió en algún siniestro. Este hecho se conoce como *riesgo de caída de cartera*, y es uno de los riesgos más relevantes en las compañías aseguradoras.

Es por tal motivo que el presente estudio tiene como objetivo predecir la probabilidad de caída de cartera de clientes en proceso de renovación para el seguro de automóviles, a partir del análisis descriptivo de cada una de las variables relacionadas con la exposición del riesgo se seleccionarán las variables que toman relevancia en el proceso de renovación de las pólizas (inicialmente para pólizas individuales) y a lo cual pueden ser el factor relevante en la causación de caída del cliente.

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

Hoy en día el tratamiento y análisis de los datos almacenados en los diversos negocios de las compañías aseguradoras, se ha convertido en una herramienta clave para el crecimiento rentable de las compañías. El análisis de datos ha tomado gran protagonismo a lo largo de los años, pues nos brinda conocimiento sobre los clientes, el negocio, los productos y las necesidades del mercado con el fin de mejorar en las propuestas y la competitividad sobre el mismo. En el sector asegurador el cálculo de las tarifas y las reservas han sido fundamentales para el crecimiento de las compañías, sin embargo, todo esto va de la mano con la gestión y los esfuerzos del área comercial. Pues dentro de las secciones del negocio para el área comercial es clave mantener unos índices de renovación de los clientes en niveles óptimos, para ello es importante mitigar las caídas de cartera en ciertos segmentos que podemos llegan a conocer que son de apetito de riesgo para la compañía.[1]

Actualmente el sector asegurador en Colombia continúa recuperándose de los efectos de la pandemia del COVID-19, los cuales a nivel mundial han golpeado de manera significativa factores macroeconómicos como lo son la inflación, la devaluación de la moneda frente al dólar y por ende el encarecimiento del costo de vida desde todos los puntos de vista personal, laboral y empresariales. Estos efectos han golpeado de manera directa el sector asegurador, en el caso del negocio de automóviles golpea directamente el costo de los siniestros pues todas las coberturas se ven afectadas por dólar e inflación directamente, como lo es el costo de los repuestos y el incremento en el costo de la mano de obra para los talleres de reparación. Lo que ha llevado desde el año 2021 a que el mercado haya tenido que garantizar su solvencia económica a partir de ajustes en las primas de los seguros lo cual garantice una competitividad en el mercado y una rentabilidad para cada una de las compañías, de la mano con el crecimiento de las compañías dentro del mismo. De esta forma, es importante para la compañía conocer a sus clientes, de manera que podamos llegar a predecir su comportamiento en el negocio, cubrir las necesidades y poder llegar a fidelizar a los clientes de forma óptima y enfocada. Este tipo de conocimiento nos brinda estrategias de retención sobre los clientes, y nos llega a generar mayor rentabilidad sobre el negocio, pues si bien se conoce que el ingreso de nuevos clientes a la cartera genera más sobre costos que el mantener la cartera fidelizada a la compañía.[2]

---

[1] Alemar Padilla, Catalina Bolance, Montserrat Guillen (2016). *Cuantificación del riesgo para la tarificación en seguros de automóvil.*

[2] Leo Guelman, Montserrat guillen (2013). *A causal inference approach to measure Price elasticity in automobile insurance.*

Teniendo en cuenta que la caída de cartera es un riesgo importante sobre el cuál se puede incurrir en una compañía aseguradora al perder la cartera. Sin embargo, es un riesgo al cuál usualmente no se le da la suficiente importancia dentro de algunas compañías de seguros. Hay que recordar que la sensibilidad de los precios dentro del mercado asegurador y especialmente en el ramo de automóviles es un tema de interés para el ramo, pues un aumento significativo en la tarifa impacta directamente a los clientes del nuevo negocio. De igual forma hay un impacto indirecto en las pólizas de renovación pues si no se posee un esquema de límites de variación de precios sobre las pólizas de acuerdo con cada uno de los perfiles de riesgo (persona – vehículo), las pólizas de renovación son aún más susceptibles a los movimientos de tarifa convirtiéndose en un mercado aún más elástico y competitivo, las cuáles históricamente en el sector asegurador suelen ser muy buenos perfiles de riesgo. Si nos centramos únicamente en la variación de prima podemos estar sesgando otras variables que nos permiten optimizar las estimaciones de precio para el negocio de renovación de las compañías, por lo cual es necesario analizar los dos subconjuntos que componen la cartera del negocio de automóviles de manera independiente y enfocado en cada una de las características propias de cada cartera respectivamente.

De esta manera, a partir de una base de datos real sobre una cartera del negocio de automóviles enfocada en vehículos livianos y en productos basados en tarifa perfil (multivariada), inicialmente tomaremos un período de tiempo de la cartera histórica para los periodos 2021 y 2022, bases con las cuales se cuenta con la información completa (todas las variables de riesgo), se realizó de manera constante el análisis del comportamiento de las renovaciones y empezamos a esperar ciertas tendencias en los índices de la renovación, las cuales por la situación del país nos sorprendieron gratamente a nivel compañía.. Donde a partir de la base de datos con todas las variables de riesgo relevantes se realizará una medición de la probabilidad de renovación de un riesgo a partir de un modelo de regresión logística (Logit) basado en un GLM. Adicionalmente se realizará una prueba con el algoritmo de Random Forest con el objetivo que este modelo sea capaz de medir el comportamiento futuro de la renovación de la cartera expuesta. Para la limpieza y transformación de la base de datos, y del cálculo de los modelos predictivos se ha utilizado el lenguaje de programación de R (RStudio) el cual como bien se conoce es un software de licencia abierta el cual nos permite modelar manera eficiente y con costos operativos de software mínimos.

2. OBJETIVOS

Objetivo General

Generar un modelo predictivo de retención de clientes para la cartera de renovación del seguro de automóviles, con el fin de direccionar acciones comerciales para la toma de decisiones de negocio, basados en información histórica de productos de negocio individual los cuales son tarifados a partir de modelos multivariables (GLM's) y centrado en grupos de vehículos livianos (automóviles, camionetas, camperos y pick ups).

Objetivos Específicos

- Determinar variables propias del riesgo que sean significativas al momento de generar la tarifa de renovación.

- Obtener un análisis descriptivo e inferencial sobre aquellas variables significativas en segmentos de interés sobre la cartera de renovación.

- Generar un modelo predictivo para obtener segmentos de interés en la cartera de renovación de automóviles.

3. MARCO TEÓRICO

Para el desarrollo del proyecto se abordará la teoría sobre dos modelos estadísticos de predicción: GLM (Modelo Lineal Generalizado) y Random Forest. Para ello se dará el análisis de la teoría existente sobre el modelo en la literatura científica y de investigación con el fin de proporcionar un conocimiento solido para la profundización de este.

Conceptos Clave

Fidelización del Cliente

La fidelización de los clientes implica la implementación de estrategias destinadas a garantizar que los clientes mantengan vínculos constantes con las compañías a lo largo del tiempo. El propósito principalmente es crear en el cliente una percepción positiva de la compañía y del negocio, lo que genere una adhesión a la compañía y/o incremente la compra de otros productos en la misma. [3]

Retención del Cliente

La retención del cliente es crucial debido a que adquirir nuevos clientes implica incurrir en un gasto mayor para la compañía. Según *Kotler (2000)* la clave para retener clientes radica en lograr su satisfacción lo cual conlleva a la fidelización del cliente. De esta forma la retención del cliente implica convertir un cliente insatisfecho en uno satisfecho y fiel a la compañía, lo cual implica llevar a cabo acciones específicas para evitar que no recompre el producto ofertado. [3]

En términos matemáticos la retención del cliente se puede medir mediante el coeficiente entre los riesgos que se renuevan y genera recaudo del pago en un periodo determinado y los riesgos candidatos a ser renovados en el mismo periodo. Este porcentaje de retención (renovación) es un indicador clave dentro de la gestión de clientes para el área comercial de las compañías, pues este nos da un factor de fidelización y nos permite llegar a entender que tipo de segmento queremos retener dentro de la compañía.

$$Retención = \frac{Número\ de\ pólizas\ renovadas\ y\ recaudadas}{Número\ de\ pólizas\ candidatas\ a\ renovar} \qquad (1)$$

[3] Sara Delfina Rosa Pierrend (2020) *La fidelización del cliente y retención del cliente: Tendencia que se exige hoy en día.*

Modelo GLM

Los modelos lineales generalizados (GLM) son una herramienta usada para modelar la relación entre una variable que busca predecir en base a un conjunto de variables independientes o explicativas. La variable predictiva o predicha se suele llamar *variable respuesta* (*target variable*) la cual denotaremos como *Y*. En la aplicación de tarifas de seguros de no vida la variable respuesta suele ser la frecuencia de los siniestros (número de siniestros sobre expuestos), el coste medio o severidad de los siniestros (valor incurrido sobre número de siniestros), la prima pura de riesgo (frecuencia por coste medio o el valor incurrido sobre los expuestos) o el loss ratio (valor de incurrido sobre el valor de prima). Usualmente para la tarificación de automóviles se suelen modelar la frecuencia y el coste medio de manera independiente, y luego si proceder a realizar el combinado entre ambos. Esto con el fin de no perder significancia de las variables si se llegase a realizar el modelo de prima pura en conjunto. [4]

De igual forma la variable respuesta puede ser la ocurrencia o no de un evento en específico, como determinar si un asegurado renovará o no su póliza o si una reclamación contiene algún tipo de fraude. En este caso el GLM se usa para estimar la probabilidad de que el evento ocurra. Para este tipo de modelos se usa más de una variable independiente para el modelo, las cuales denotaremos como $x_1, \dots, x_i$ donde *i* será el número de variables independientes o predictores del modelo, las cuales suelen ser el tipo de vehículo, la edad, la marca, el departamento, la experiencia siniestral del cliente, etc. Así la función sería la siguiente:

$$Y = \beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i + \varepsilon \qquad (2)$$

En esta función el parámetro β_0 representa el valor del intercepto cuando las variables dependientes son 0, lo coeficientes $\beta_1, \dots, \beta_i$ se estiman mediante el modelo del GLM. Para poder calcular el predictor lineal debemos transformar mediante una función de enlace la variable de respuesta, la cual nos permitirá mayor flexibilidad para relacionar la predicción del modelo con los predictores del mismo, es decir, nos permite construir un modelo el cuál se ajustará mejor a las necesidades del proyecto. En este caso la usará la función de enlace Logit ya que nos permite acotar los valores de predictores entre 0 y 1, la transformación lineal se realizará a

[4] Mark Goldburd, Anand Khare, Dan Tevet, Dmitry Guller (2019) *Generalized Linear Models for Insurance Rating.*

partir de la función sigmoide link y el método de estimación de Máxima Verosimilitud.

$$Función\ Sigmoide = \frac{1}{1 + e^{-x}} \qquad (3)$$

Si se sustituye el valor de la *Y* (variable respuesta) por la función anterior, se realiza la modificación a la función derivando y aplicando la inversa del logaritmo natural en base a funciones exponenciales en ambos lados de la ecuación tenemos:[5]

$$P\ (Y = 1) = \frac{e^{\beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i}}{1 + e^{\beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i}}$$

$$\ln\left(\frac{P\ (Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i$$

Denotaremos como $\mu_i = \frac{P\ (Y=1)}{1-P(Y=1)}$ así:

$$\mu_i = e^{\beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i} = e^{\beta_0} * e^{\beta_1 * x_1} * \ldots * e^{\beta_i * x_i} \qquad (4)$$

Los modelos multiplicativos o multivariables suelen ser los más usados en la tarificación de los seguros de no vida, dado que son modelos:

- Sencillos y prácticos de aplicar
- La naturaleza intuitiva de los modelos nos permite aumentar o decrecer las primas en un porcentaje dependiendo del peso de la variable dentro del mismo, más no el hecho de sumar factores fijos por defecto. Lo cual permite cobrar primas con un mejor ajuste de rentabilidad y enfocado a la suficiencia del ramo.
- Siguiendo el principio de parsimonia nos permite tener mejores estimaciones con una mayor simplicidad en el proceso.

Ahora bien, hay que tener presente los diferentes tipos de variable respuesta y la función de enlace a usar, a continuación, se dará un resumen de las más habituales para los modelos paramétricos usados. [6]

[5] Francisco Folk (2022) *Modelización de la probabilidad de no renovación para una cartera de Salud.*

[6] Arroyo Indira, Bravo Luis, Llinas Humberto, Muñoz Fabian (2014) *Distribuciones Poisson y Gamma: Una discreta y Continua Relación*

Distribución Poisson

Su función de enlace se encuentra mediante un *log link,* donde el valor esperado de la variable respuesta se representa como $E(Y_i) = \mu_i$ donde μ_i es la frecuencia de ocurrencia del evento esperado, y la función de enlace usada es

$$g(\mu_i) = \log \mu_i \qquad (5)$$

donde se observa que el objetivo del modelo es la frecuencia de ocurrencia.

Distribución Gamma

Su función de enlace se encuentra mediante un *link reciproco* donde se denota como

$$g(\mu_i) = \frac{1}{\mu_i} \qquad (6)$$

como bien se sabe se usa para modelar variables aleatorias positivas mediante un proceso de Poisson, y logra medir el tiempo transcurrido hasta obtener la ocurrencia del evento esperado. En el negocio de automóviles la distribución Gamma se usa para el modelamiento de la severidad o coste medio de las ocurrencias.

Distribución Binomial

Su función de enlace se encuentra mediante *logit link* donde el valor esperado de la variable respuesta se representa como $E(Y_i) = n\pi_i$ donde π_i representa la probabilidad de éxito, y su función de link se representa como

$$g(\mu_i) = log\left(\frac{\mu_i}{n- \mu_i}\right) \qquad (7)$$

sí reemplazamos $\mu_i$tenemos

$$g(\mu_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) \qquad (8)$$

Así se evidencia que el objetivo de nuestro modelo es una función que nos arroja la probabilidad de éxito u ocurrencia de nuestra variable respuesta.

Esta será la distribución para usar dentro del modelo con el fin de obtener la probabilidad de renovación de los riesgos mediante un modelo de GLM.

Random Forest

El Random Forest es un modelo estadístico empleado para realizar predicciones y clasificaciones de datos, creado por Leo Breiman entre 1996 y 2002. El Random Forest crea un conjunto de árboles de decisión usando aleatoriamente un porcentaje de los datos y las variables de cada modelo, hasta alcanzar un nodo de finalización para el mismo. Luego se realiza un promedio de todas las predicciones realizadas en cada uno de los árboles generando así una predicción final.

En el Random Forest todos los árboles tienen la misma distribución dentro del bosque completo, sin embargo, son diseñados de tal forma que sean distintos entre sí, lo que ayuda a reducir la correlación de estos. Se crearon combinando la técnica de Bagging (*Bootstrap Aggregation)* de Breiman (1996) y la selección aleatoria de atributos en la base de datos. El objetivo principalmente es mejorar la varianza de Bagging y por ende la correlación entre los árboles. La diferencia entre el Bagging y el *Bootstrapping* se centra en la selección de la base de datos distinta (cada árbol es distinto entre sí). [7]

En un resumen la generación de los árboles se realiza de la siguiente manera:

1. Se divide de manera aleatoria la base de datos entre train y test.
2. Se genera el bosque aleatorio en la base de train y cada árbol se construye con el siguiente proceso:
    a. Se seleccionan *n* datos realizando Bootstrap (repetición) de la base de train.
    b. Esta muestra es usada para entrenar la cantidad de árboles *i*.
    c. Si tenemos *M* cantidad de inputs, se selecciona una cantidad *m* de ellos para usarse en la decisión de cada nodo del árbol. Este valor *m* se mantiene para todo el bosque.
    d. Se iteran los valores para cada input *m* con particiones de acuerdo con los criterios de los hiperparametros.
    e. Se itera el proceso varias veces hasta la ejecución de todos los *i* árboles.
3. Se realiza el promedio de todas las predicciones hasta obtener el análisis final.

Así los hiperparametros principales son la cantidad de árboles usados, el número *M* de predictores usados en cada uno de los árboles y el tamaño mínimo de los nodos

[7] Salomón Micael, Carranza Juan, Piumetto Mario, Monzani Federico, Montenegro Marzo, Cordoba Augusto (2018) *Random Forest como técnica de valuación masiva del valor del suelo urbano: una aplicación para la ciudad del río cuarto, Cordoba, Argentina.*

para cada uno. El objetivo de estos hiperparametros es optimizar el modelo de tal manera que este se ejecute en mejores tiempos y con mayor exactitud en los resultados de significancia del modelo. Estos hiperparametros se prueban varias veces registrando el error en cada uno de ellos y eligiendo los mejores para el modelo, teniendo en cuenta si el valor de los hiperparametros no es elegido correctamente, esto muchas veces puede arrojar mayor error al modelo en test.

Fundamentos para el Análisis de Resultados

Con el fin de analizar las predicciones de los modelos inicialmente obtendremos una matriz de confusión y ciertas mediciones estadísticas que de acuerdo con su valor nos darán la explicación de la exactitud y demás parámetros relevantes del modelo.

Matriz de Confusión

La matriz de confusión nos muestra los valores reales (columnas) y los valores de referencia o predichos (filas), quedando una tabla de la siguiente forma:[8]

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	Verdaderos Negativos	Falsos Negativos
1	Falsos Positivos	Verdaderos Positivos

Tabla 1 Ejemplo Matriz de Confusión

A partir de esta matriz se crean las mediciones estadísticas tales como: exactitud, precisión sensibilidad y especificidad, así como otras métricas adicionales necesarias para el análisis de un GLM:[9]

> ***NOTA:*** *Los 4 ítems de la matriz se denotarán por sus iniciales para mayor facilidad en las fórmulas. Ejemplo: Verdaderos Negativos se notará por VN, y Falsos Positivos por FP.*

[8] La matriz de confusión y sus métricas https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[9] https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e

Accuracy (Exactitud)

Es el porcentaje de verdaderos acertados en el modelo total, aquí se suman los *verdaderos positivos* y los *verdaderos negativos* sobre el total de los registros analizados.

$$Accuracy = \frac{VP + VN}{VP + FN + FP + VN} \qquad (9)$$

Precisión (Precision)

Es el porcentaje de verdaderos positivos sobre el total de positivos.

$$Precision = \frac{VP}{VP + FP} \qquad (10)$$

Sensibilidad (Sensitivity)

Es el porcentaje de verdaderos positivos acertados en el modelo.

$$Sensitivity = \frac{VP}{VP + FN} \qquad (11)$$

Especificidad (Specificity)

Es el porcentaje de verdaderos negativos acertados en el modelo.

$$Specificity = \frac{VN}{FP + VN} \qquad (12)$$

Tasa de Verdaderos Positivos

Es la razón de verdaderos positivos sobre el total de verdaderos reales.

$$Verdaderos\ Positivos = \frac{VP}{VP + FN} \qquad (13)$$

Tasa de Verdaderos Negativos

Es la razón de verdaderos negativos sobre el total de negativos reales.

$$Verdaderos\ Negativos = \frac{VN}{FP + VN} \qquad (14)$$

Kappa

Es el dato que mide la concordancia de los datos considerando los valores esperados y predichos del modelo (concordancia observada). Es usada usualmente cuando el modelo está basado en una distribución binomial. Ya que suele tener en cuenta la ocurrencia por azar, su primera medición se le atribuye a Galton (1892) y Smeeton (1985).

Inicialmente se calcula el *accuracy* del modelo, para sí calcular la probabilidad de que el resultado sea por azar a partir de la probabilidad de acierto o error tanto para los valores reales como los predichos, de tal forma que se tiene:

$$P(Predichos = 0) = \frac{VP + FP}{VP + FN + FP + VN} \quad (15)$$

$$P(Predichos = 1) = \frac{FN + VN}{VP + FN + FP + VN} \quad (16)$$

$$P(Reales = 0) = \frac{VP + FN}{VP + FN + FP + VN} \quad (17)$$

$$P(Reales = 1) = \frac{FP + VN}{VP + FN + FP + VN} \quad (18)$$

Dado esto se hace la probabilidad esperada dada la probabilidad condicional de cada estado:

$$P(Esperada) = (P(Predichos = 0) * P(Reales = 0)) + (P(Predichos = 1) * P(Reales = 1)) \quad (19)$$

Así tenemos:

$$Kappa = \frac{P(Observada) - P(Esperada)}{1 - P(Esperada)} \quad (20)$$

Es necesario tener en cuenta que para todas las medidas estadísticas anteriormente mencionadas si tenemos valores superiores a 0.8 esto significa una buena medición estadística de acuerdo con su significado. [10]

[10] La matriz de confusión y sus métricas https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

4. METODOLOGÍA

Para el desarrollo del análisis se realiza la creación de la base de datos a partir de la base de pólizas candidatas a renovar, la base de pólizas vigentes a diferentes cortes de tiempo y bases de anulación donde podemos identificar si la póliza fue anulada completamente durante el periodo de tiempo. Una vez creada la base de datos definitiva se procede a hacer el análisis de las características propias de la población de cartera vigente y candidata a renovar para los periodos de enero 2021 y Diciembre 2022 del seguro de automóviles en negocios individuales tarifados por modelos multivariables en MAPFRE Seguros de Colombia S.A, dentro de la cual se tienen todas las variables del riesgo que se usan para realizar la tarifa y variables adicionales las cuales se va a analizar su significancia dentro de los modelos.

Limpieza de Datos

La base fue construida con toda la información que se cuenta dentro de la compañía, se tienen en cuenta la base de pólizas candidatas a renovar, la base de riesgos vigentes mes por mes y bases de pólizas anuladas, a partir de la unión de esta información se crea la base de riesgos renovados validando que cada uno de los índices por negocio coincidan con la información real dentro de la compañía.

Inicialmente se contaba con la base entre 2021 y 2022, con un total de 220.796 riesgos candidatos a renovar. Se procede a realizar la limpieza de los datos, como homologar variables para tener categorías únicas dentro de cada variable, crear rangos para algunas variables categóricas o numéricas muy extensas como lo son el score financiero, la antigüedad del vehículo, la edad del asegurado, las fechas de las bases de renovación. Se valida la base de datos en los campos vacíos (NA's) algunos datos son completados con la categoría base de referencia, otros son depurados de la base de datos para evitar distorsiones sobre la misma. Se realizan filtros sobre la información con el fin de solamente obtener riesgos de vehículos livianos y que pertenezcan a productos individuales de tarifa perfil. Adicionalmente se revisan los riesgos atípicos, en base a primas muy altas fuera de los rangos de desviación de la tasa media de la cartera de automóviles. Toda esta información es de la compañía por lo cuál las bases no pueden ser mostradas dentro del proceso académico.

Finalmente después del proceso de depuración interna de las bases de datos, el cuál abarco más del 50% aportado al desarrollo del proyecto, de un total de 220.796 riesgos nos quedamos con la información base de 202.238 riesgos correspondientes a un 91.59% del total de la información inicialmente recolectada.

Análisis de Datos

En la siguiente tabla se muestran las variables que se van a manejar en la base de datos junto con su significado y el tipo de dato que este posee. Para tener en claro la variable respuesta o el *target variable* la **MCA_RENOV** la cual nos indicará si finalmente un riesgo fue o no renovado en la base histórica a modelar. Algunas variables no fueron incorporadas dentro de los modelos, como fecha de inicio y fin de vigencia, fechas de nacimiento y valor de primas tanto anterior como de renovación, sin embargo, las fechas de inicio y fin de vigencia de las pólizas fueron usadas para contemplar la exposición de los riesgos, teniendo en cuenta que la mayoría de riesgos en el seguro de automóviles son de exposición 1 (vigencia completa) para el caso de esta base de datos histórica dado que son riesgos en cartera candidatos a continuar en la compañía. Para los valores de prima, se establece un *rango de prima anterior* el cuál ha sido determinado por el esquema de suavización del área de pricing de la compañía.

TIPO DE VARIABLE	VARIABLE	DEFINICIÓN	PREDICTIVA
Categórica	**PRODUCTO**	Producto por el cuál se emite la poliza, hay productos individuales, programas de marca y de más. Van desde el producto 110 hasta el 133	NO
Numérica	**NUM_POLIZA**	Numero de póliza, identificador para la compañía del riesgo	NO
Fecha	**FECHA_INICIO**	Fecha de inicio de vigencia del reisgo candidato a renovar	NO
Fecha	**FECHA_FIN**	Fecha de fin de vigencia del reisgo candidato a renovar	NO
Numérica	**ID**	Número de identificación del asegurado	NO
Categórica	**REGIONAL**	Estructura comercial de la compañía en el país	SI
Categórica	**SUCURSAL**	Subconjunto de la regional donde se emite el riesgo	NO
Categórica	**OFICINA**	Subconjunto de la sucursal donde se emite el riesgo	NO
Categórica	**AGENTE**	Identificación del agente comercial que emite el riesgo	NO
Categórica	**FASECOLDA**	Código fasecolda, identificador de la categoría, marca y línea del vehículo.	NO
Categórica	**TIPO_VEHICULO**	Tipo de vehiculo distribuido en Automoviles (1), Camionetas (2) y Pick Ups (3)	SI
Numérica	**MODELO**	Año de la versión de venta del vehiculo	NO
Categórica	**MARCA**	Marca del vehiculo	SI
Categórica	**LINEA**	Referencia del vehiculo a más detalle	NO
Categórica	**CHASIS**	ID único de la carrocería del vehiculo	NO
Categórica	**PLACA**	ID único del vehiculo	NO
Numérica	**VA_COMERCIAL**	Valor asegurado Comercial, sobre el cuál se asegura y tarifica el riesgo en las coberturas de Pérdidas Parciales	NO
Numérica	**VA_NUEVO**	Valor asegurado a Nuevo, sobre el cuál se asegura y tarifica el riesgo en las coberturas de Pérdidas Totales	NO
Numérica	**RC**	Monto Límite de Cobertura para la cobertura de Responsabilidad Civil	NO
Categórica	**FIDELIZACION**	Categoría asignada comercialmente al asegurado, de acuerdo a la antigüedad en la compañía, la cantidad de pólizas en los diversos ramos de la misma.	SI
Numérica	**SINIESTROS**	Cantidad de siniestros en la vigencia inmediatamente anterior a la postulación de las pólizas candidatas.	SI
Numérica	**BONUS_MALUS**	Descuento técnico del asegurado, asignado de acuerdo a su experiencia siniestral históricamente en el mercado asegurador Colombiano en pólizas todo riesgo.	SI
Numérica	**IMP_DESC_COMERCIAL**	Descuento comercial otorgado a la vigencia inmediatamente anterior de parte del intermediario al momento de emitir la póliza.	NO
Numérica	**PRIMA_ANTERIOR**	Prima Neta emitida en la vigencia inmediatamente anterior a la candidata a renovación.	NO
Numérica	**COMISION_ANTERIOR**	Comisión correspondiente a la vigencia inmediatamente anterior a la candidata a renovación.	NO
Numérica	**PRIMA_PRERENOV**	Prima Neta propuesta a la renovación, teniendo en cuenta los topes de suavización asignados.	NO
Numérica	**COMISION_PRERENOV**	Importe de Comisión correspondiente a la prima Neta propuesta de renovación.	NO
Numérica	**FEC_RENOV**	Llave de Año-Mes del periodo de inicio de la renovación	NO
Categórica	**LLAVE**	Llave de Chasis_Placa_Fecha renovacion	NO
Categórica	**NEGOCIO**	Tipo de negocio sobre el cuál se emite, Individual o programas de marca (Chevy, Renault o Toyota)	SI
Numérica	**ANTIGÜEDAD**	Antigüedad del vehiculo de acuerdo al Modelo y Año de la Renovación	SI
Categórica	**RANGO_ANTIGUEDAD**	Rangos de Antigüedad del vehiculo	NO
Categórica	**COD_CIUDAD**	Codigo Dane de la ciudad del riesgo	NO
Categórica	**DEPARTAMENTO**	Departamento del riesgo emitido	SI
Fecha	**FEC_NACIMIENTO**	Fecha de Nacimiento del Asegurado	NO
Numérica	**EDAD**	Edad del asegurado a la altura de renovación	NO
Binaria	**MCA_VIG**	Marca 0 si la póliza NO es encontrada en riesgos vigentes (a corte de mes) / 1 si la póliza es encontrada en riesgos vigentes (a corte de mes)	NO
Categórica	**CLASE_SPTO**	Si la póliza posee movimiento de Anulación en el periodo de vigencia correspondiente. Pueden ser Anulaciones Completas (el riegso no genera ni un día de exposición en la compañía) o Anulaciones Parciales (el riesgo genera al menos un día de exposición en la compañía y genera la devolución de la prima no devengada).	NO
Fecha	**FEC_ANULACION**	Fecha desde la que se genera la Anulación del riesgo.	NO
Numérica	**IMP_PRIMA_ANUL**	Importe de la prima anulada, valor de prima no devengada a devolver al cliente.	NO
Numérica	**ALTURA_ANULACION**	Diferencia entre la Fecha de Anulación y la Fecha de Inicio de Vigencia del Riesgo, dividiendo en 365 días para dar el dato en años.	NO
Binaria	**MCA_ANUL**	Marca 1 si el riesgo NO generá suplemento de anulación / 0 si el riesgo generá suplemento de anulación	NO
Binaria	**MCA_RENOV**	Es el producto entre la MCA_VIG y MCA_ANUL, lo cuál generá 0 si el riesgo no fue renovado y 1 si generó renovación.	TARGET
Fecha	**FECHA_EMISION**	Fecha Inicial del primer suplemento de emisión de la póliza.	NO
Numérica	**ALTURA_RENOV**	Diferencia entre la Fecha de Inicio de Vigencia y la Fecha de emisión, dividido en 365 días para dar el dato en años entero.	SI
Categórica	**RANGO_PRIMA_ANT**	Categorización por rangos de la PRIMA_ANTERIOR	SI
Categórica	**TIPO_CAJA**	Tipo de Caja del vehiculo, manual, automático, triptonico o desconocido.	NO
Numérica	**PESO_POTENCIA**	Relación entre el peso del vehiculo y la potencia que este posee.	NO
Categórica	**RANGO_PESO_POTENCIA**	Clasificación interna de Peso_Potencia la cuál nos indica que tan rapido o lento es un vehiculo de acuerdo con sus caracteristicas mecanicas.	NO
Categórica	**TIPO_PERSONA**	Si es persona Natural o NIT.	NO
Categórica	**SCORE**	Calificación financiera de la persona, dato recolectado por fuentes externas lo cuál indica el nivel de riegso crediticio de la persona.	NO
Numérica	**SUAVIZACION**	Diferencia entre la Prima Emblem y los topes máximos o mínimos de las bandas de renovación en la compañía.	NO
Numérica	**PRIMA_EMBLEM**	Prima cotizada como riesgo nuevo a partir del cotizador vigente en cada altura del tiempo.	NO
Numérica	**VAR_PRIMA**	Variación de la prima de renovación frente a la prima anterior que pago el cliente. Es decir, el incremento de valor de prima al momento de la renovación.	SI
Categórica	**RANGO_SCORE**	Clasificación del SCORE por rango caracteristicos.	NO
Categórica	**RANGO_EDAD**	Clasificación de la EDAD por rango caracteristicos.	NO
Categórica	**GENERO**	Hombre, Mujer, NIT o Sin Información.	SI
Categórica	**RANGO_VAR_PRIMA**	Clasificación de la VAR_PRIMA por rangos con los que se suavizan las tarifas actualmente.	NO

Tabla 2: Diccionario de Variables

Análisis Descriptivo de la Población

Inicialmente se realizará el análisis de la distribución de las variables con el fin de limpiar y realizar ajustes sobre la base de datos. Para ello se usarán diagramas de Pareto el cual nos mostrará la frecuencia acumulada del número de riesgos en cada categoría por variable, de igual forma en un gráfico y tabla paralela se mostrarán los índices de renovación de las mismas categorías y variables.

Se procede a observar el ***tipo de negocio*** en los cuales se observa la concentración de más del 70% de los datos sobre el negocio individual, los otros 3 negocios restantes son convenios de marca que maneja la compañía, los cuales se tarifican y se renuevan de la misma manera que un producto individual, por lo cual se decide mantener este tipo de negocio dentro de la base de datos. Adicionalmente se observa que el negocio Individual posee un % de renovación (retención) del 80.4%, seguido de Toyota con un 76.2% y Renault y Chevrolet con un 70% aproximadamente.

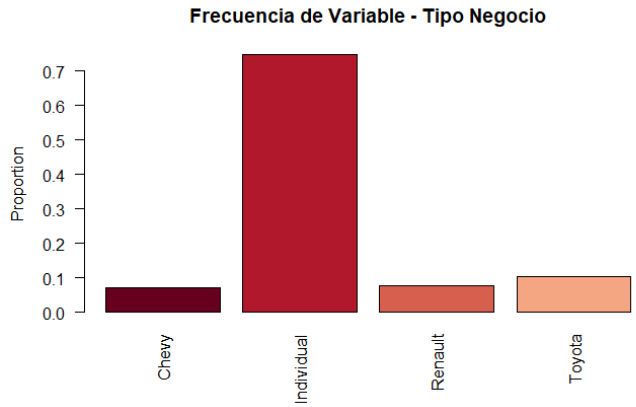

Ilustración 1 Distribución de Tipo de Negocio

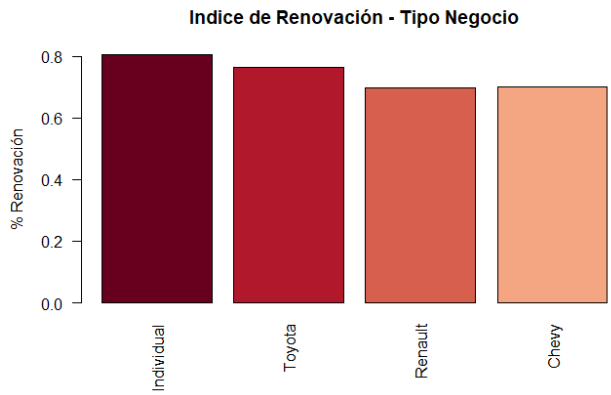

Ilustración 2 Índice de Renovación Tipo de Negocio

TIPO NEGOCIO	Riesgos Renovados	Total Candidatas	% Renovación
Individual	121.432	151.024	80,41%
Toyota	16.104	21.132	76,21%
Renault	10.877	15.602	69,72%
Chevy	10.159	14.480	70,16%
TOTAL	158.572	202.238	78,41%

Tabla 3 Índice de Renovación Tipo de Negocio

Se observa el *Producto* el cuál es la descripción más a detalle del tipo de negocio. Se observa una correlación entre estas dos variables, pues el producto 118 el cual tiene más del 40% de los datos corresponde a los productos Individuales, mientras el 126 es Chevy, 131 es Renault y 131 es Toyota, los demás son del tipo de negocio individual. En general se observan los índices de renovación entre el 70% y el 80% para todos los productos.

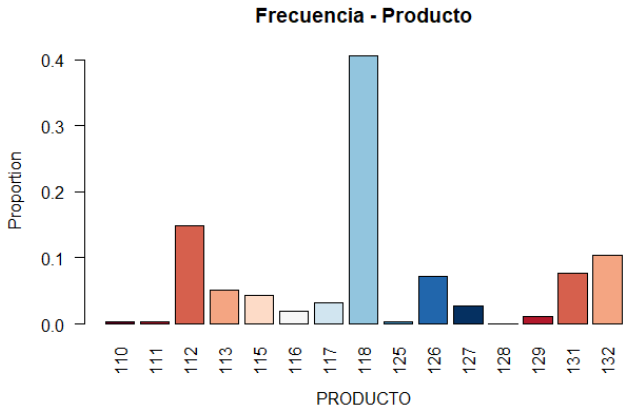

Ilustración 3 Distribución Producto

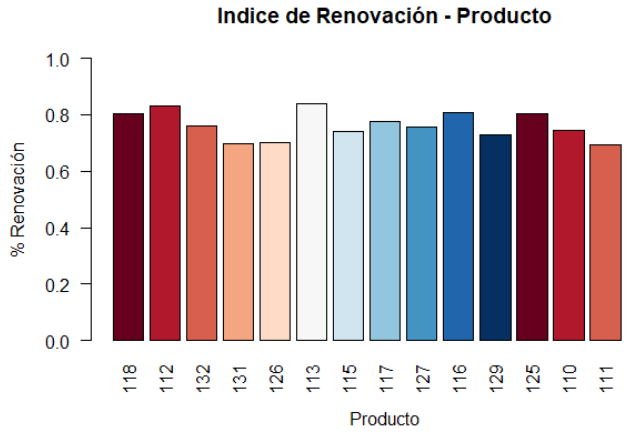

Ilustración 4 Índice de renovación del Producto

Para la fecha de renovación se observa una gran cantidad de riesgos expuestos en enero, abril y diciembre del 2021 y 2022 con temporalidades iguales. Esto se debe a comportamientos de altos índices de ventas durante años anteriores, estos meses se han observado con cierta estacionalidad alta en ventas para la compañía. De igual forma, se observa para los meses de junio de 2021 y 2022 que su exposición en riesgos es muy baja, esto se debe a que durante el año 2020 por la crisis de la Pandemia COVID-19 en Colombia se genera la Circular Externa 021 de 2020[11] por medio de la cual la Superintendencia Financiera de Colombia sobre la cual las compañías aseguradoras están en la obligación de devolver 1 mes de prima no devengada a todos los clientes o generar la extensión correspondiente a 1 mes de vigencia de la póliza, con el fin de mitigar el impactos para los asegurados dadas las restricciones generadas por la pandemia. De la misma forma para los meses de junio se observa que dado el mes de extensión de vigencia y el efecto de la crisis global en índice de renovación para este periodo de tiempo es ligeramente inferior a la media del restante de meses en la base de datos.

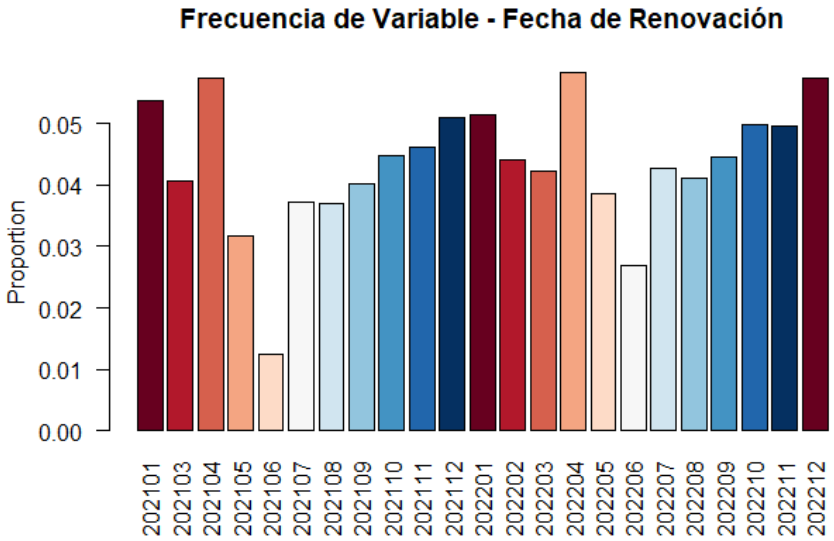

Ilustración 5 Distribución Fecha de Renovación

[11] https://www.superfinanciera.gov.co/descargas/institucional/pubFile1046025/ce021_20.docx

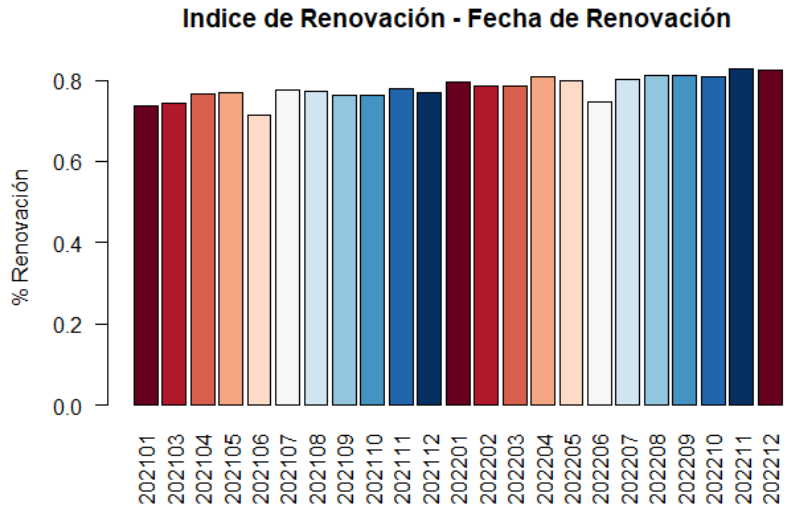

Ilustración 6 Índice de Renovación para la Fecha de Renovación

Para la variable de ***marca*** dado que se muestran varias marcas en un solo gráfico se complementa la información con el detalle de las primeras 10 marcas las cuales abarcan el 90% del total de la base de datos. Se observa claramente que para las 3 principales marcas (Chevrolet, Renault y Toyota) se da un incremento en los índices de renovación al ver toda la base de datos compacta y no abierta a nivel de negocio, de las demás marcas se ve en términos generales un índice y renovación entre el 78% y 80% constante.

MARCA	Riesgos Renovados	Total Candidatas	% Renovación
Chevrolet	31.295	41.253	75,86%
Renault	30.195	39.601	76,25%
Toyota	25.973	32.903	78,94%
Ford	10.526	13.272	79,31%
Nissan	10.899	13.184	82,67%
Mazda	10.154	12.669	80,15%
Kia	9.721	12.420	78,27%
Volkswagen	5.862	7.206	81,35%
Hyundai	5.333	6.590	80,93%
Suzuki	2.621	3.336	78,57%
TOTAL	142.579	182.434	78,15%

Tabla 4 Índice de Renovación por Marca

Para el ***tipo de vehículo*** se observa que más del 90% de la base se concentra en los grupos 1 y 2, automóviles y camionetas respectivamente, mientras el grupo 3 pick ups solo tiene menos del 10% de la base de datos, sin embargo, en términos de la frecuencia no se observa gran diferencia entre ellos, el grupo 1 y 3 tienen frecuencia del 77% mientras el grupo 2 frecuencia del 79%.

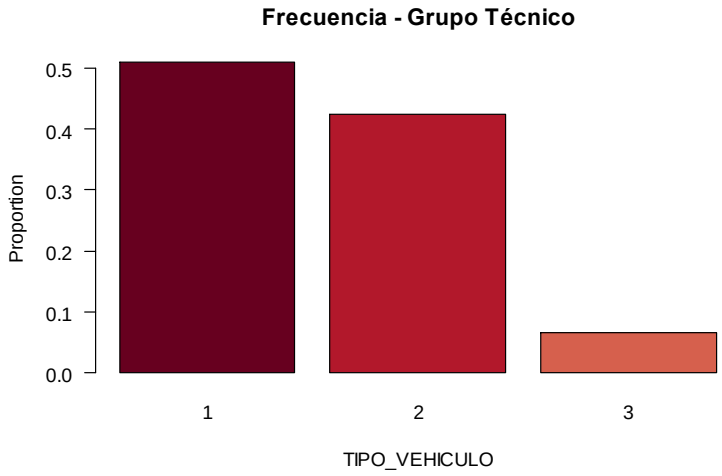

Ilustración 7 Distribución por Tipo de vehículo

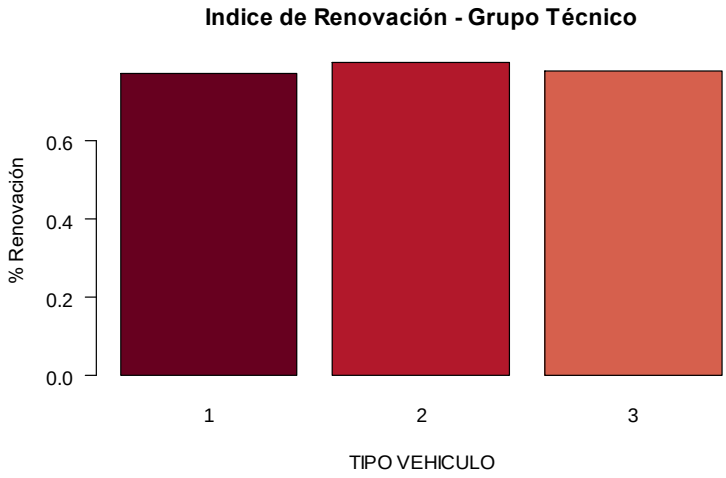

Ilustración 8 Índice de Renovación por Tipo de vehículo

Para la ***regional*** se observa que la mayor concentración de expuestos está en regionales de Bogotá, Dirección Comercial y Norte Centro. El índice de renovación no varía mucho entre ellos, a excepción de los negocios de corredores, sin embargo, su peso sobre la base de datos es mínimo, lo cual no es significativo al momento de desviar información en los modelos.

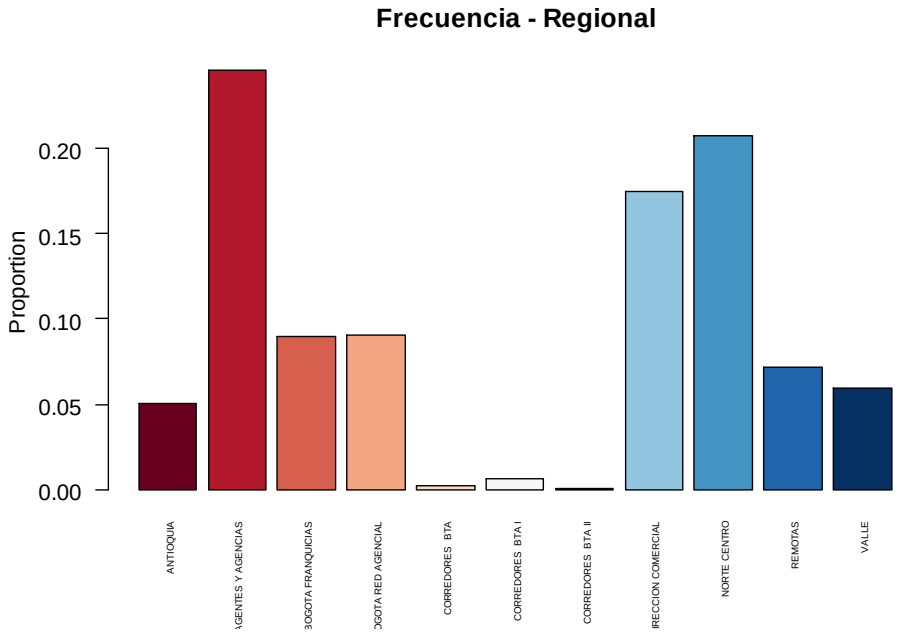

Ilustración 9 Distribución por Regional

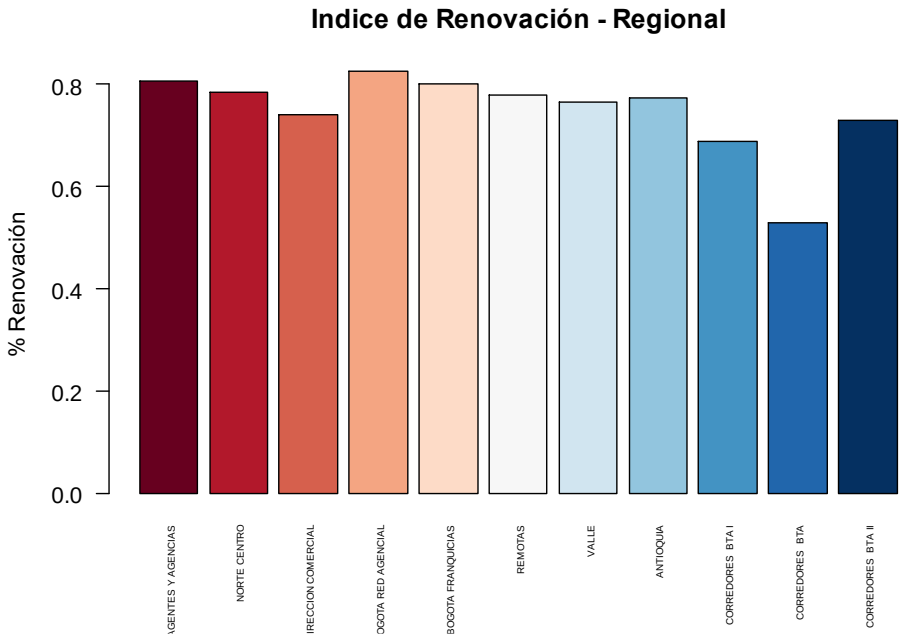

Ilustración 10 Índice de Renovación por Regional

Para la ***cantidad de siniestros*** durante la vigencia anterior, no hay masa significativa para valores mayores a 1, por lo cual se pretende no usar la variable ya que no muestra significancia en la distribución de esta. Y los índices de renovación se observan disminuyendo a medida que tiene más cantidad de siniestros, esto dado el castigo que hay en los límites de la variación de prima al momento de la renovación.

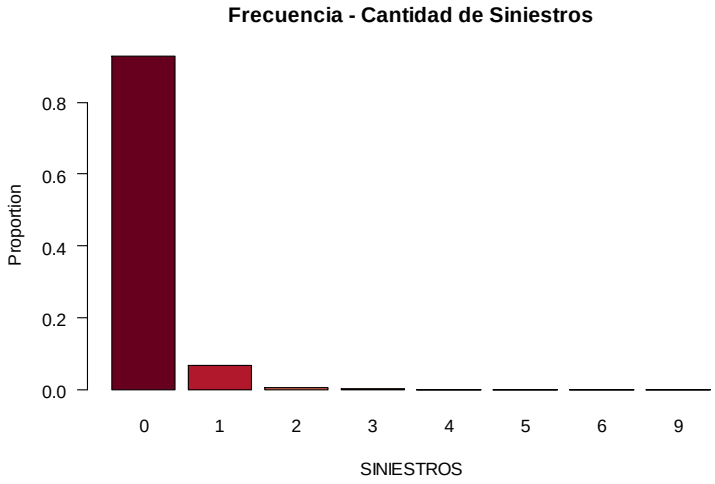

Ilustración 11 Distribución de Cantidad de Siniestros en vigencia anterior

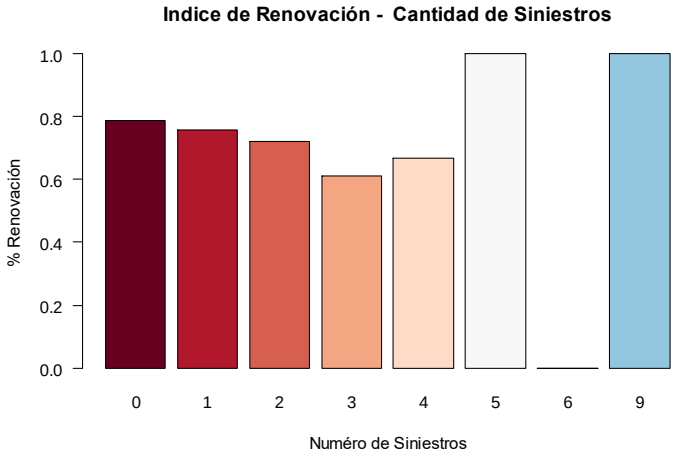

Ilustración 12 Índice de renovación para la Cantidad de Siniestros en vigencia anterior

En el caso de ***Años sin Siniestro o Bonus Malus*** se observa una gran concentración en el Bonus de 4 con más de un 60% de la información, seguido por Bonus de 1 cercano al 20% de la información, Bonus de 2 y 3 cercano al 10% de la información y por último bonus de 0 con un mínimo en la renovación. Respecto a los índices de renovación de esta variable se observa que a medida que la calificación siniestral de cliente mejora los índices de renovación son más altos, con el caso del bonus 4 un índice del 82%. Lo cual es un comportamiento esperado en el seguro de automóviles.

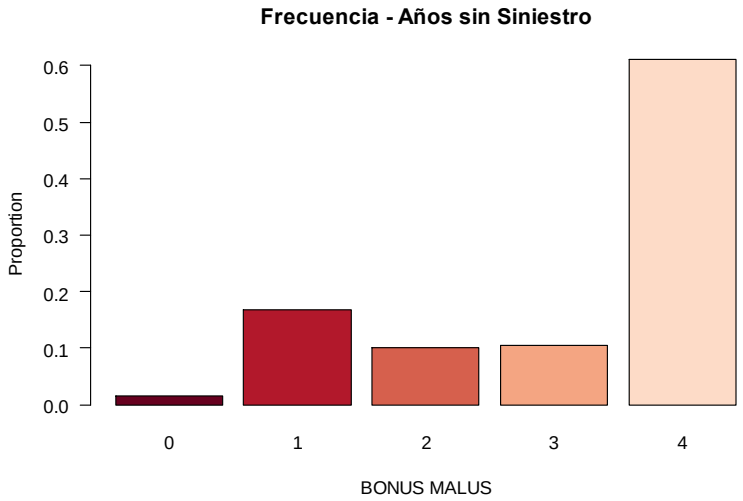

Ilustración 13 Distribución de los Años Sin Siniestro

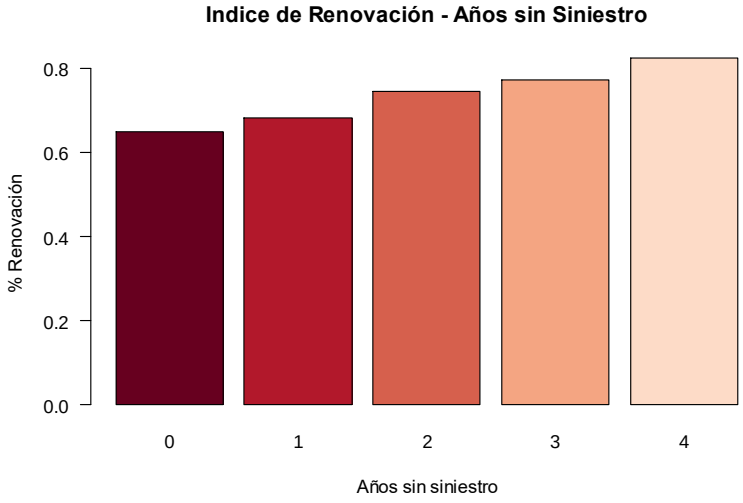

Ilustración 14 Índice de Renovación de los Años sin Siniestro

Para la ***altura de renovación*** se observa que la distribución de esta variable está más cargada a la izquierda, es decir, el 50% de riesgos a renovar son riesgos que ingresaron a la compañía por primera vez en su vigencia inmediatamente anterior y la curva va disminuyendo de acuerdo con que la altura va aumentando. Respecto al índice de renovación para la altura 1 siendo el 50% de la masa tiene un índice de renovación del 73% bastante inferior a la media observada del 78%, y a medida que la altura de renovación aumenta la fidelización del cliente aumenta respectivamente llegando a niveles del 84% de renovación para clientes de más de 5 años en la compañía y de más del 90% de renovación para más de 9 años de antigüedad en la compañía.

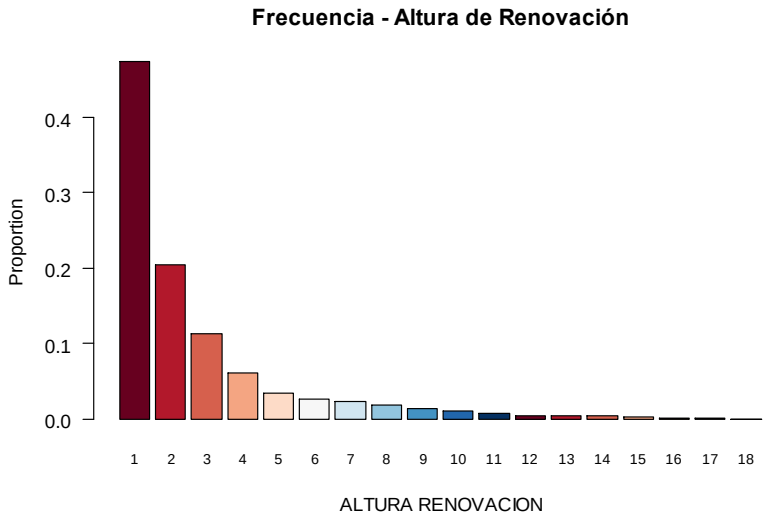

Ilustración 15 Distribución Altura de Renovación

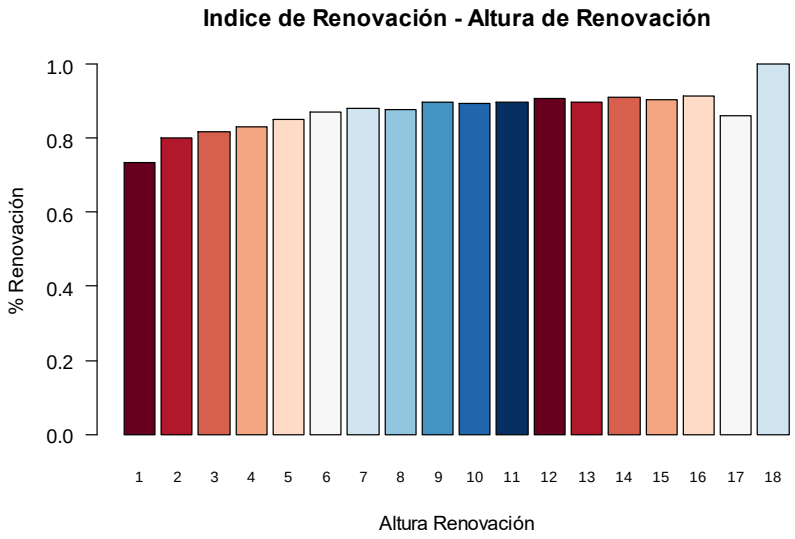

Ilustración 16 Índice de Renovación por la Altura de Renovación

Para la ***categoría de fidelización*** se observa la mayor concentración en la categoría Plata y Sin Fidelización los cuales en su mayoría pertenecen a los negocios del programa de marca Renault, Chevy y Toyota. La categoría Oro y Platino dentro de la compañía son riesgos sobre los cuales se aplica un descuento adicional, por el hecho de llevar muchos años dentro de la compañía, tener más de 1 producto en diferentes negocios dentro de la compañía y otras características las cuales son analizadas desde el área comercial. Así mismo se observa que estas dos categorías Oro y Platino tienen índices de renovación del 87% promedio.

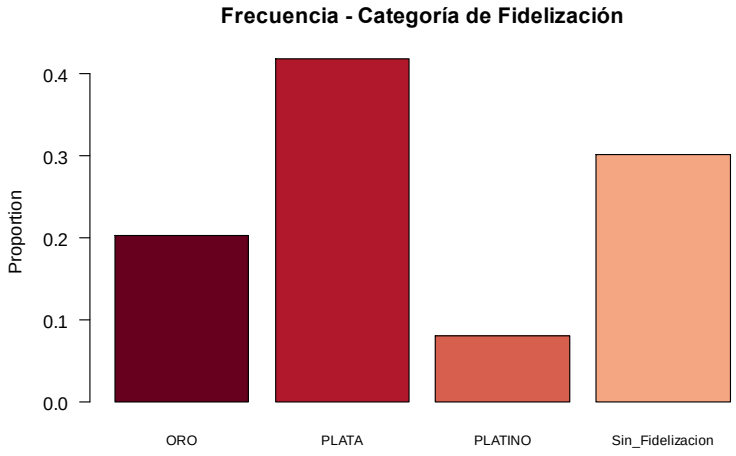

Ilustración 17 Distribución Categoría de Fidelización

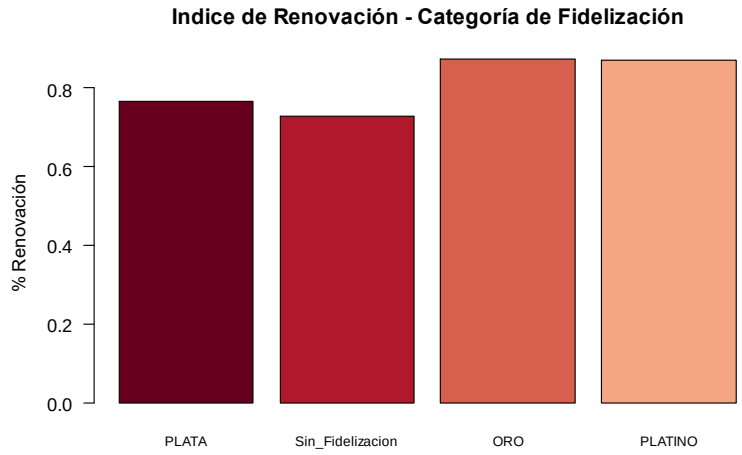

Ilustración 18 Índice de Renovación Categoría de Fidelización

Para el ***límite de RC*** se observa que la mayoría de los riesgos poseen una RC límite de 3000 seguido de una de 2000, sin embargo, para los índices de renovación no se observa gran diferencia entre ellos, hay un incremento para las RC de 500 y 750 con índices del 85% sin embargo también son riesgos que por sus características pueden tener una diferencia significativa en primas para el mercado.

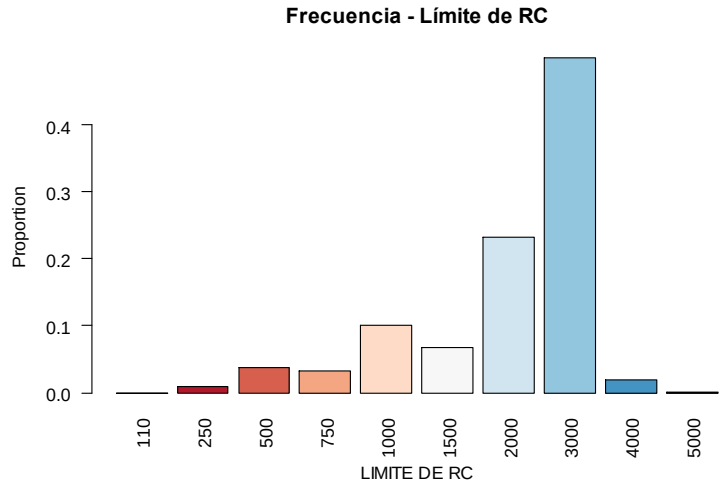

Ilustración 19 Distribución Límite RC

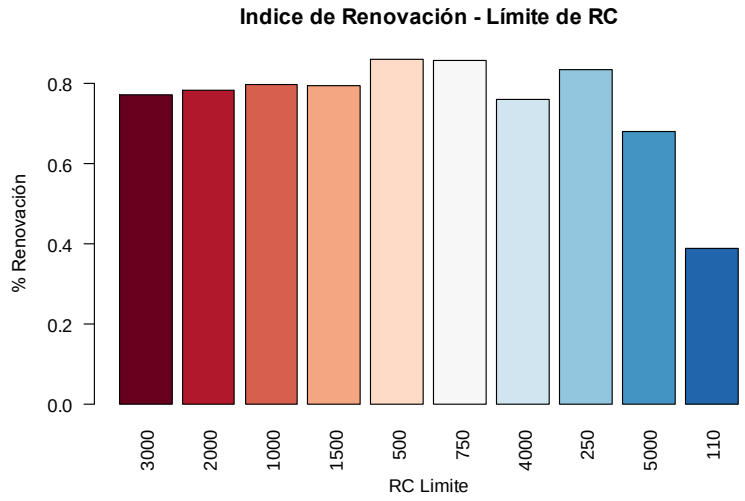

Ilustración 20 Índice de Renovación por Limite de RC

Para el ***rango de prima anterior*** se observa el 40% de la información en rangos entre 1 millón y 1,5 millones de pesos (COP), seguido del 30% por primas mayores a 1,5 millones de pesos. Respecto al índice de renovación se observa que a menor valor de prima anterior mayor es el índice de renovación.

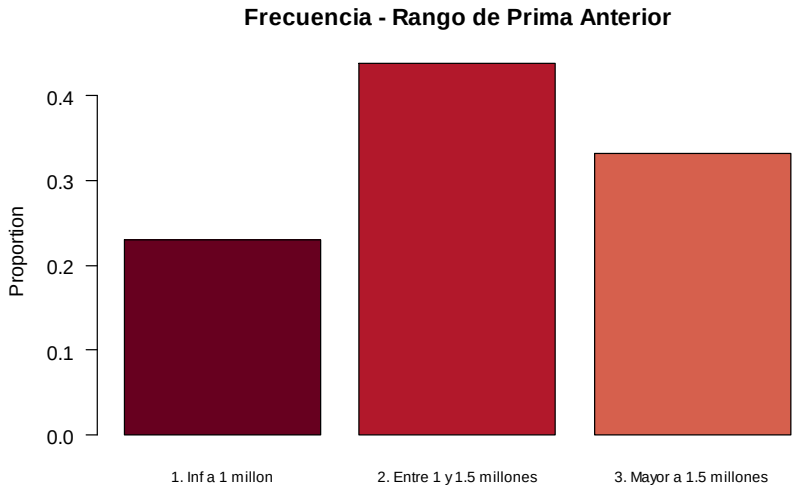

Ilustración 21 Distribución Rango de Prima Anterior

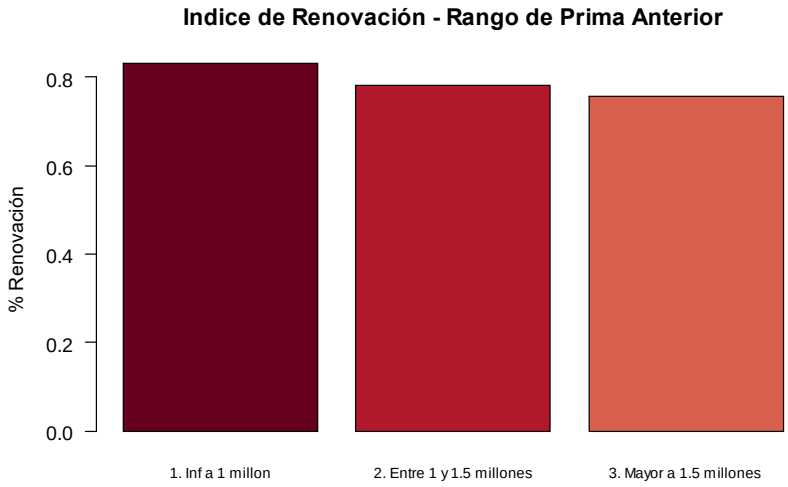

Ilustración 22 Índice de Renovación por Rango de Prima Anterior

Para el ***rango de antigüedad de vehículo*** se observa la concentración de vehículos con antigüedades entre 1 y 10 años, esto teniendo en cuenta que el parque asegurado en Colombia sabemos que desde hace unos años se empezó a concentrar en estas antigüedades de vehículos. Respecto al índice de renovación se observa que a medida que el vehículo es más antiguo mayor es el índice de renovación, lo cual puede tener alguna correlación con la altura de renovación de los clientes.

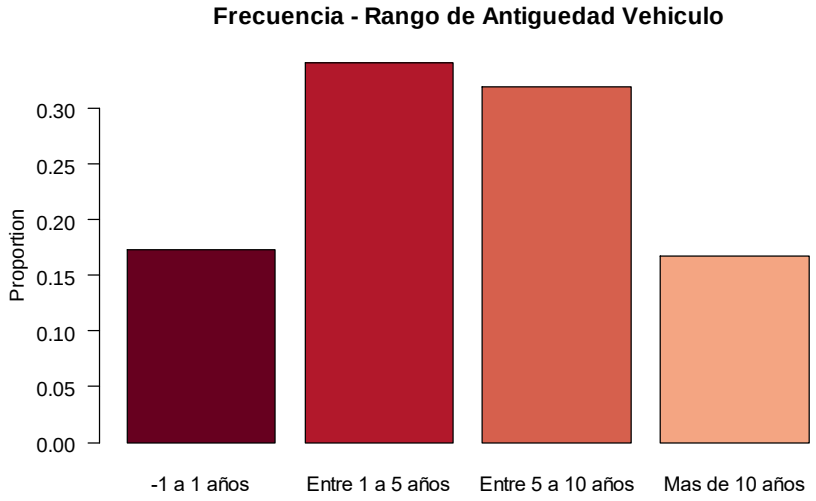

Ilustración 23 Distribución por Rango de Antigüedad del Vehículo

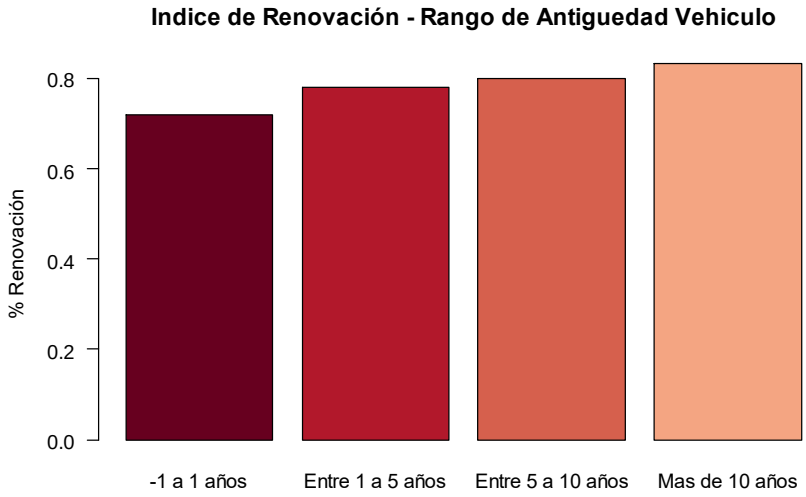

Ilustración 24 Índice de Renovación por Antigüedad del Vehículo

Para la ***edad*** de los asegurados se observa claramente una distribución normal de la cartera, reflejando claramente una campana de Gauss sobre la misma. Respecto al índice de renovación se observa una ligera tendencia de incremento a medida que la edad va aumentando en los clientes.

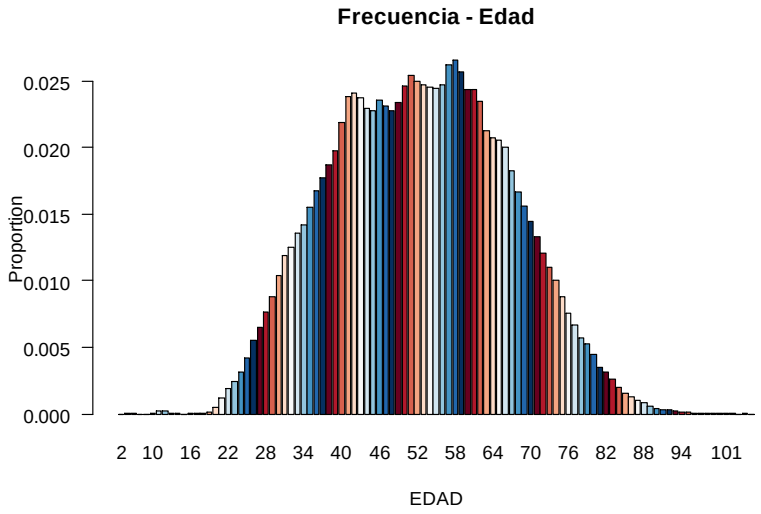

Ilustración 25 Distribución Edad

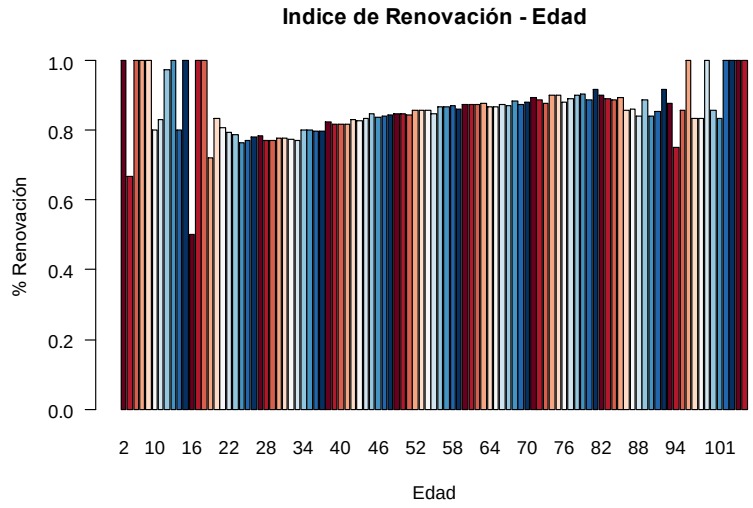

Ilustración 26 Índice de Renovación por Edad

Si realizamos agrupaciones para los rangos de edad, podemos confirmar la afirmación anterior, a medida que la edad del cliente aumenta este renueva con un mayor índice en la compañía.

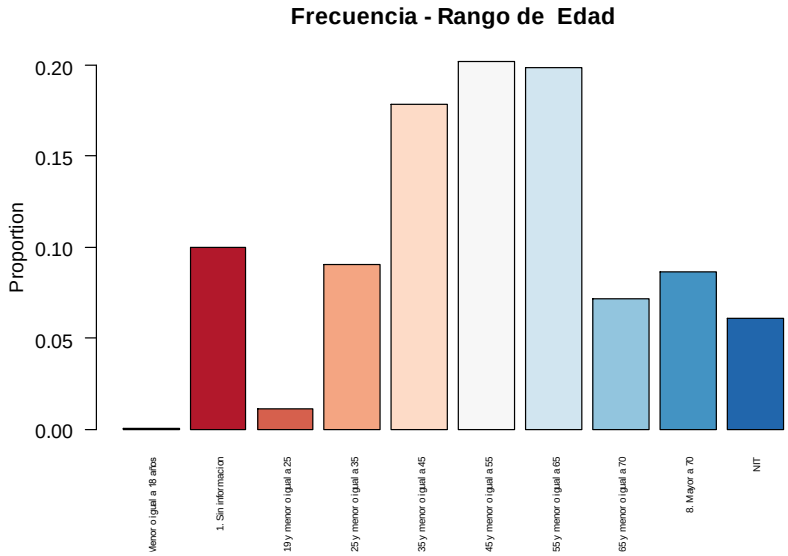

Ilustración 27 Distribución Rango de Edad

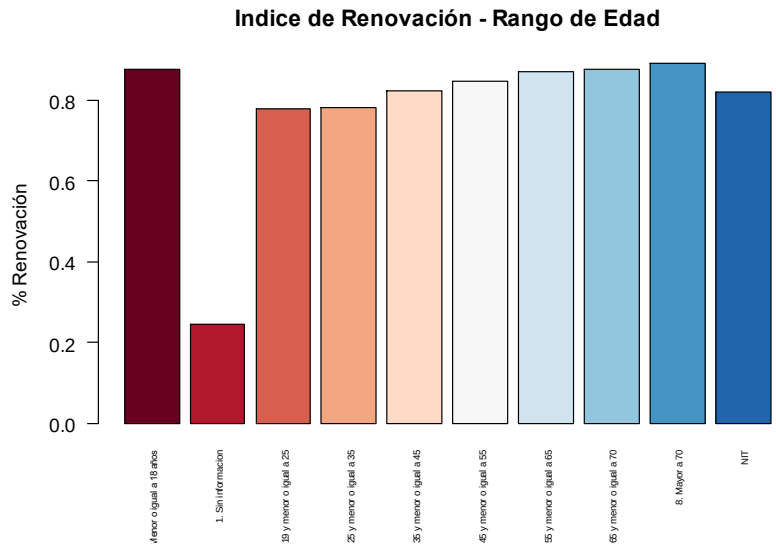

Ilustración 28 Índice de renovación Rango de edad

Para el ***género*** se observa la mayor concentración en persona natural. Respecto al índice de renovación, se puede observar que aquellos riesgos que no poseen género en el sistema no renuevan existiendo una correlación directa entre ellos.

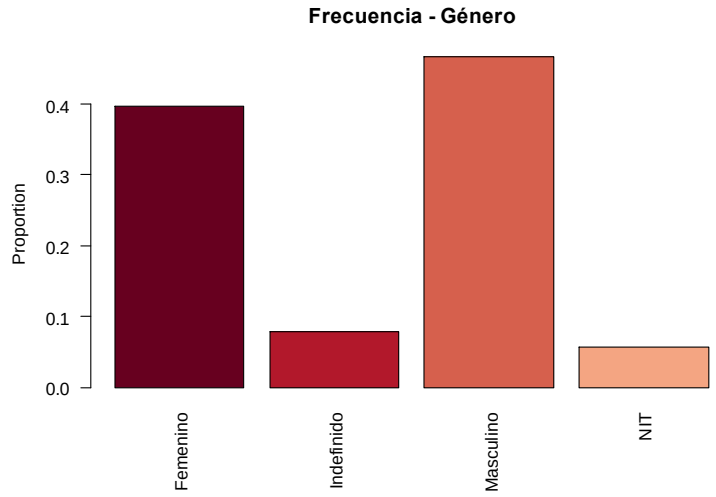

Ilustración 29 Distribución por Género

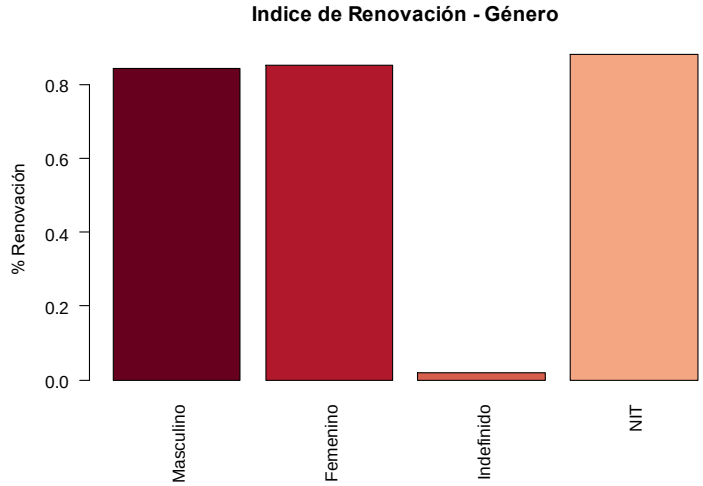

Ilustración 30 Índice de Renovación por Genero

Para la ***variación de prima*** respecto a la prima anterior se observa que la distribución está en todos los rangos de prima, pero la mayor masa la podemos observar en variaciones del 0% al 18%, sin embargo, respecto al índice de renovación se puede observar que, dado el entorno atípico de variables macroeconómicas, para variaciones por encima del 0% el índice de renovación se mantiene cercano al 80%.

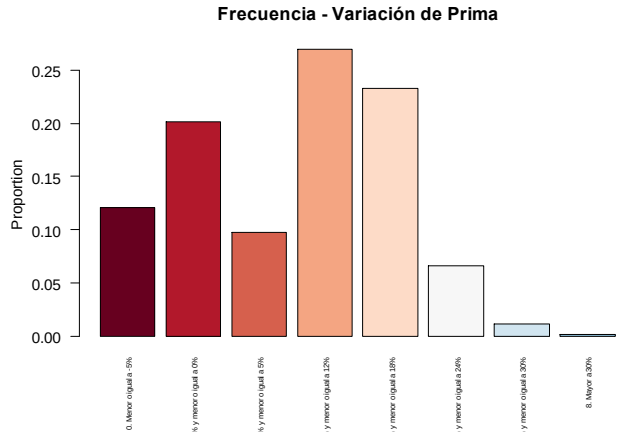

Ilustración 31 Distribución por Variación de Prima

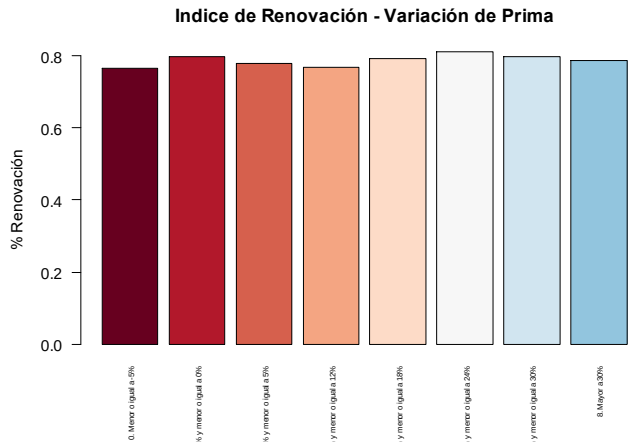

Ilustración 32 Índice de Renovación por Variación de prima

Respecto al ***tipo de persona*** se puede observar que toda la masa está en persona natural lo cual se encuentra bastante correlacionado con el género, respecto al índice de renovación no se observan diferencias.

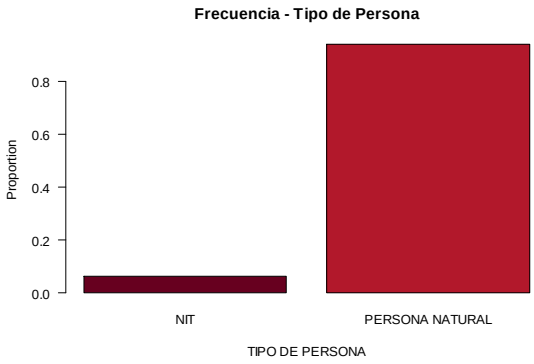

Ilustración 33 Distribución por Tipo de Persona

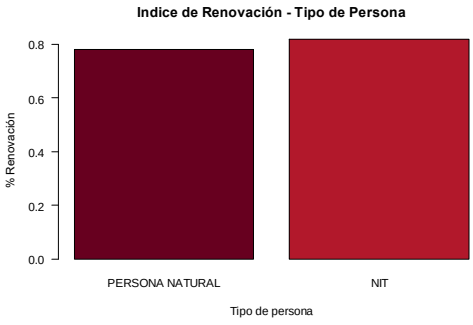

Ilustración 34 Índice de Renovación por Tipo de Persona

Respecto al ***rango del score financiero*** se observa la masa de 750 a 900 puntos de score financiero y se observa que a mayor score financiero se observa un mayor índice de renovación.

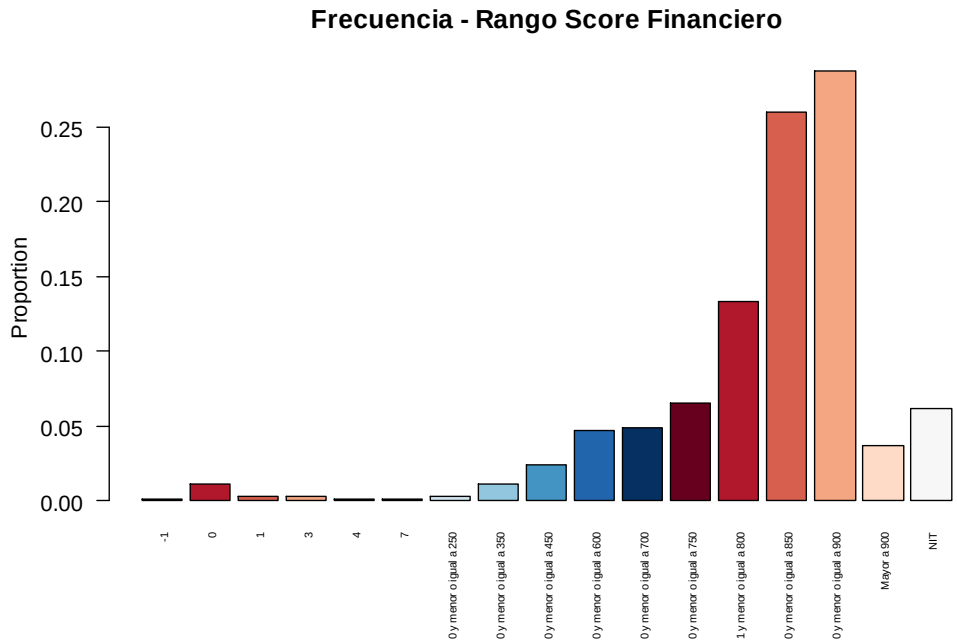

Ilustración 35 Distribución Rango Score

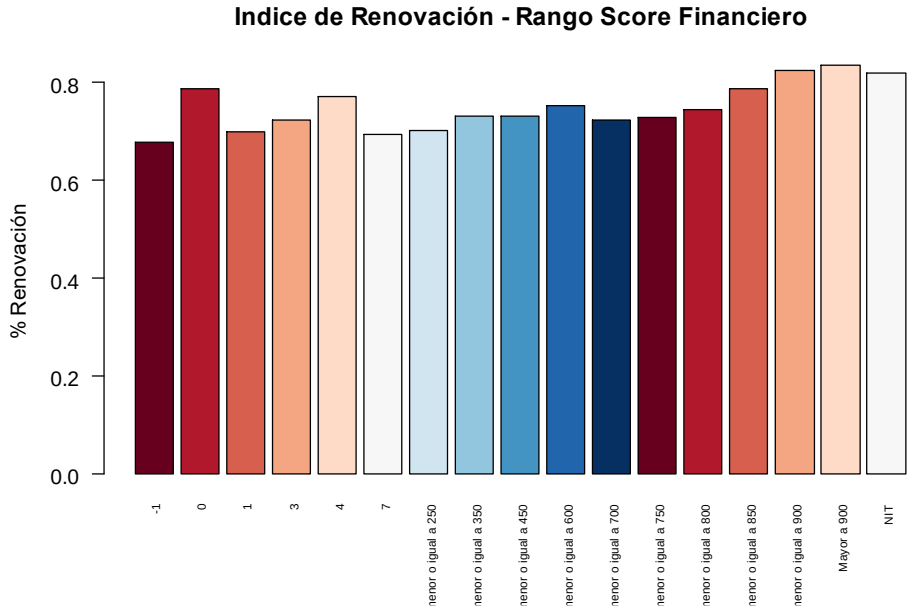

Ilustración 36 Índice de Renovación por Rango de Score

Para el ***rango de peso potencia*** se observa que la base de datos enfoca al 80% de los riesgos en un rango Normal, es decir, la relación de peso/potencia del vehículo no se tiene cartera de vehículos tipo deportivos en la cartera. El índice de renovación es inherente al rango.

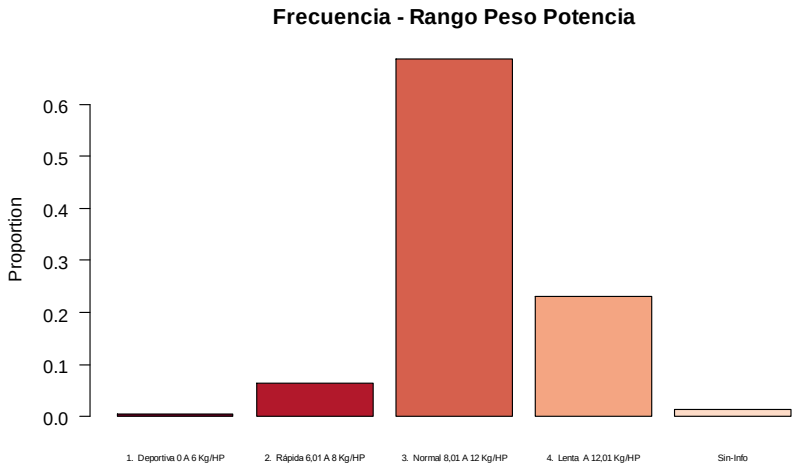

Ilustración 37 Distribución por Peso Potencia

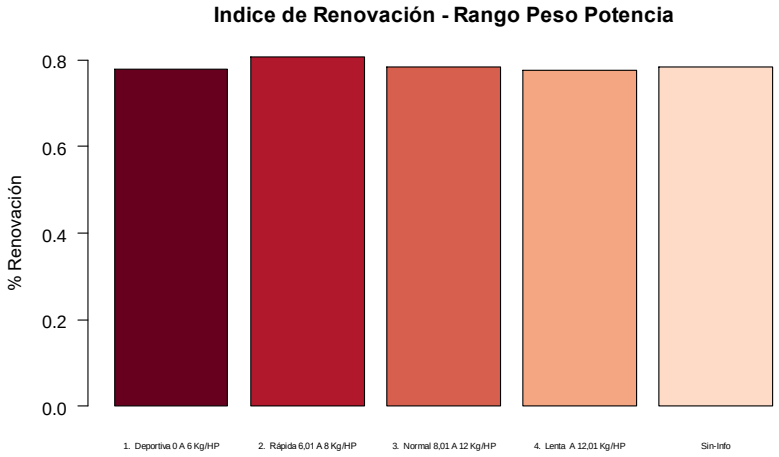

Ilustración 38 Índice de Renovación por Peso Potencia

Para el ***tipo de caja*** se observa que el parque observado en su mayoría es tipo MT (mecánico) y TP (triptónico), sin embargo, referente a los índices de renovación no hay variación por tipo caja, en el tipo caja AT (automático) se incrementa el índice porque puede llegar a tener una correlación con el tipo de vehículo 2 que son camionetas, las cuales en su mayoría son Automáticas.

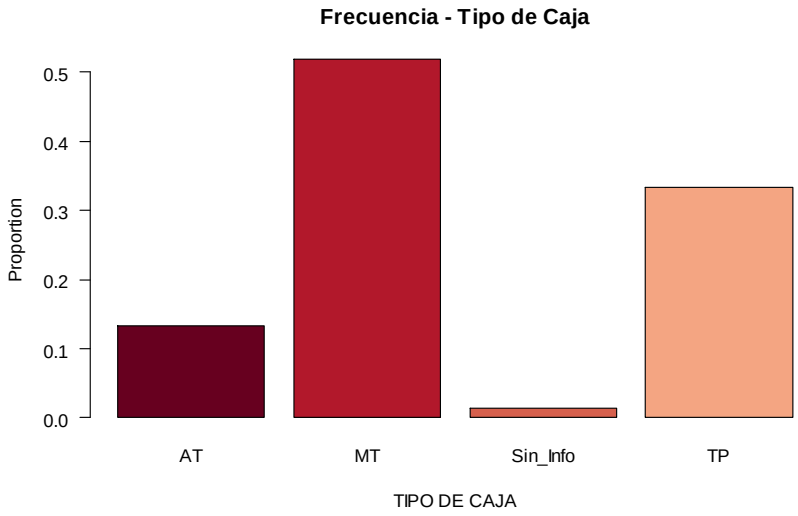

Ilustración 39 Distribución Tipo Caja

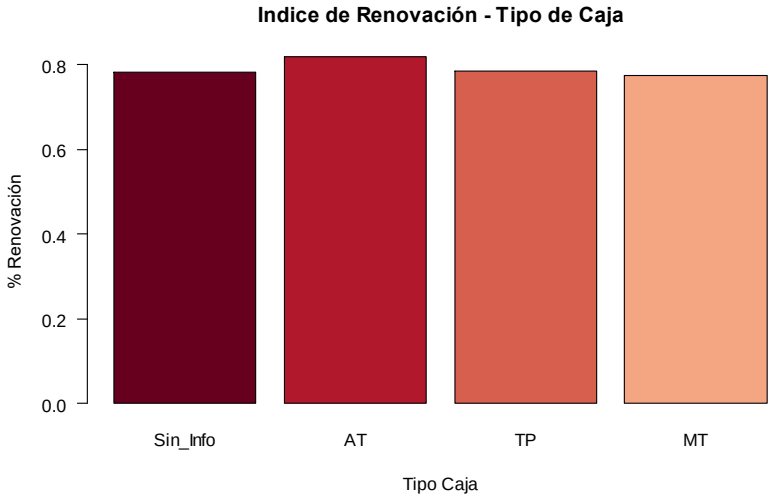

Ilustración 40 Índice de Renovación por tipo Caja

De esta forma se concluye con el análisis descriptivo de cada una de las variables, lo cual ya nos da apertura a poder seleccionar ciertas variables que pueden ser significativas en nuestro modelo GLM.

Correlación entre Variables del Modelo

La correlación entre las variables es un aspecto clave para los modelos de líneas, ya que la correlación entre variables puede afectar la estimación de los coeficientes y su significancia dentro del mismo, además el hecho de mantener variables correlacionadas dentro del modelo va en contra del principio fundamental de modelación: la parsimonia. [12]

Para el cálculo de la correlación se abre la base en dos grandes subconjuntos, uno de variables categóricas sobre el cuál realizamos análisis de correlación por medio de V-Cramer y otro de variables numéricas donde analizamos por medio del coeficiente de correlación de Pearson. En ambos casos el principio de análisis de correlación se mantiene, siendo variables muy correlacionadas todas aquellas que tengan valores cercanos a ± 1.0, y para una mayor exactitud tomaremos en este caso valores superiores a ± 0.7. [13]

La siguiente imagen muestra la matriz de correlación para las variables categóricas:

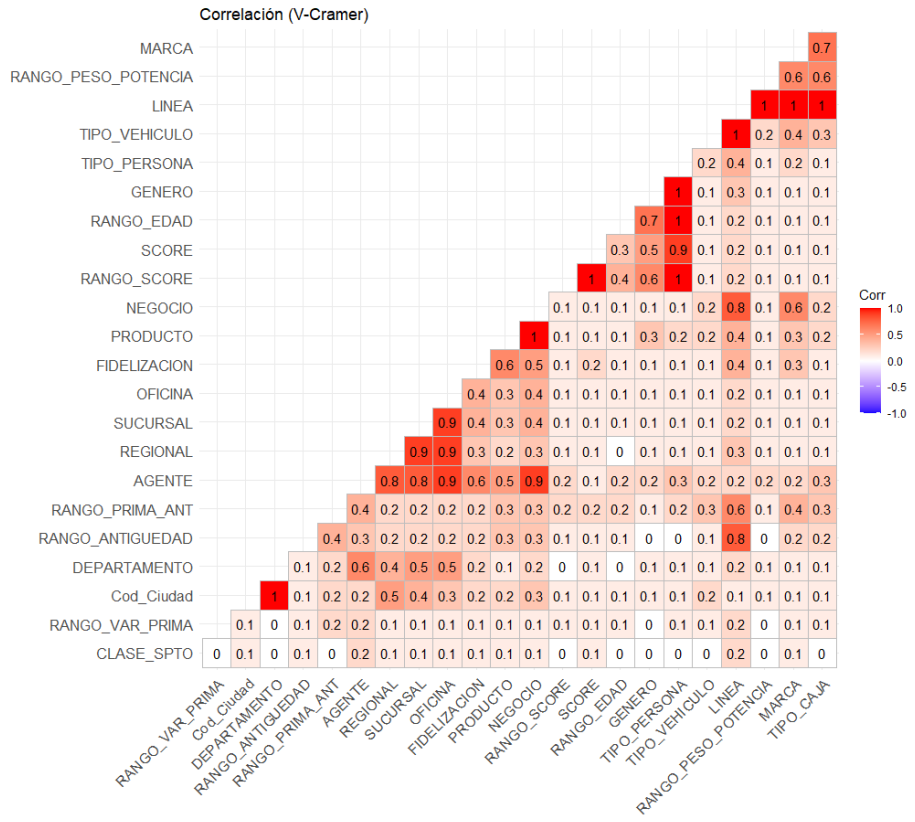

Tabla 5 Correlación V-Cramer (Categóricas)

[12] García, J. L., Chagolla, H., & Noriega, S. (2015). *Modelos: efectos de la colinealidad en el modelado de regresión y su solución. Cultura Científica y Tecnológica*

[13] Juicio de Experto, uso general.

A partir de lo cual podemos determinar que las variables que presentan mayor correlación para el modelo son:

- Línea
- Tipo Persona
- Score
- Producto
- Sucursal
- Regional
- Agente
- Tipo Caja
- Rango Peso Potencia
- Código Ciudad

Para las variables numéricas se muestra la siguiente matriz a partir de coeficientes de Pearson:

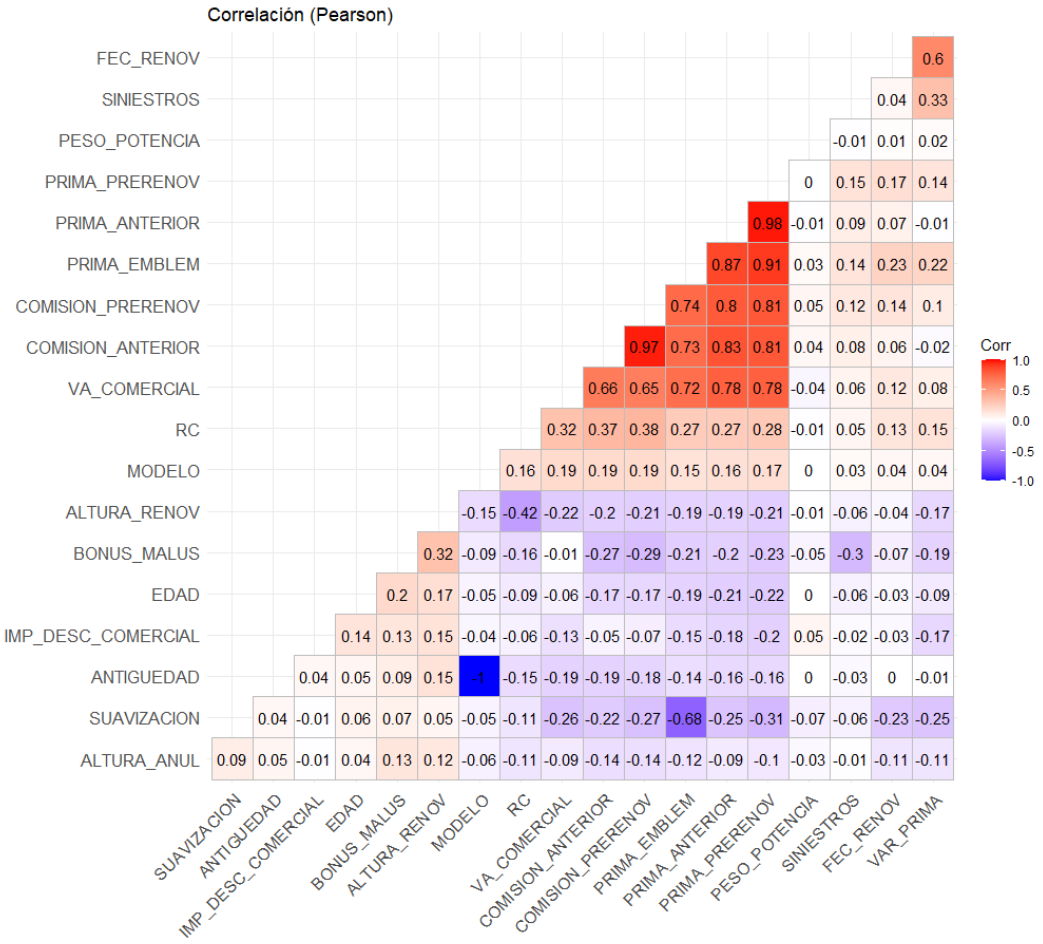

Tabla 6 Correlación Pearson (Numéricas)

Donde se observa que las variables con mayor correlación para el modelo son:

- Prima Anterior
- Prima Prerenovación
- Prima Emblem
- Comisión Prerenovación
- Comisión Anterior
- Comisión Anterior
- Valor comercial
- Modelo

A partir del análisis realizado se retiran las variables que muestran correlaciones altas y se vuelve a analizar la matriz de correlaciones con el fin de validar los coeficientes en la nueva data.

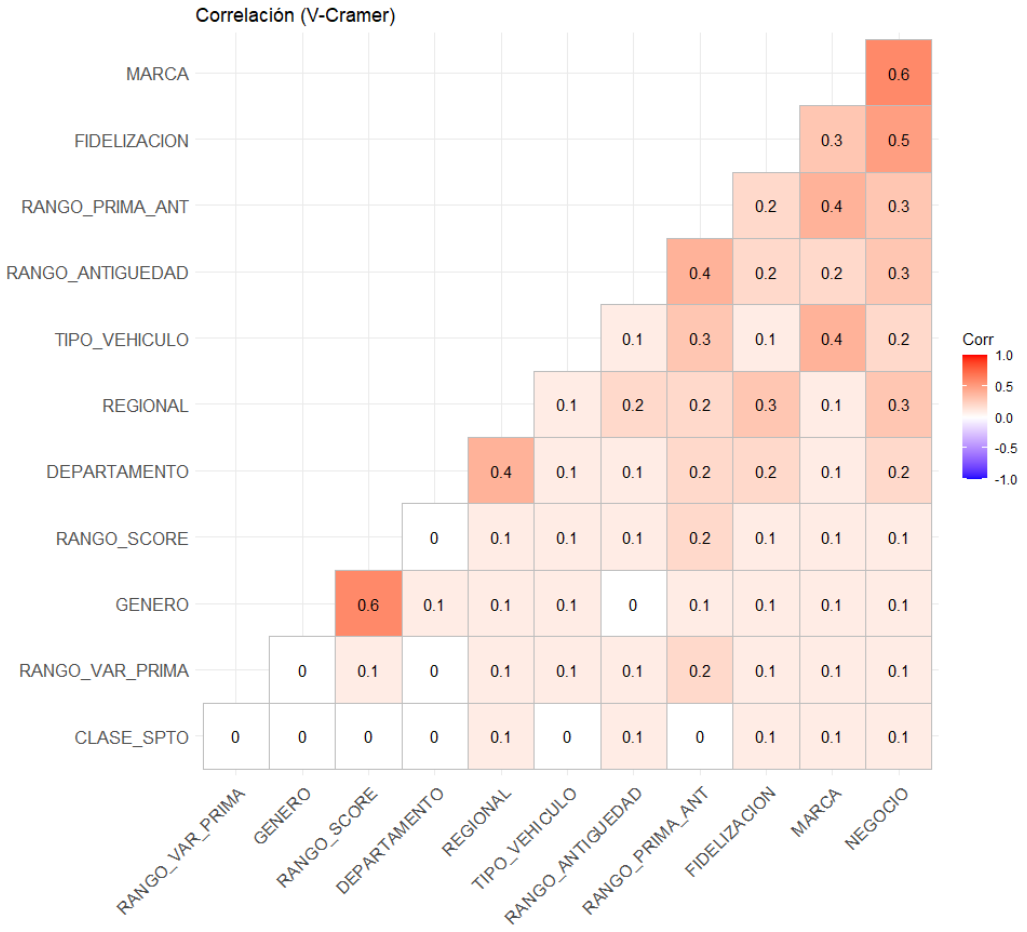

Tabla 7 Correlación V-Cramer (Categóricas) BBDD Sin correlaciones

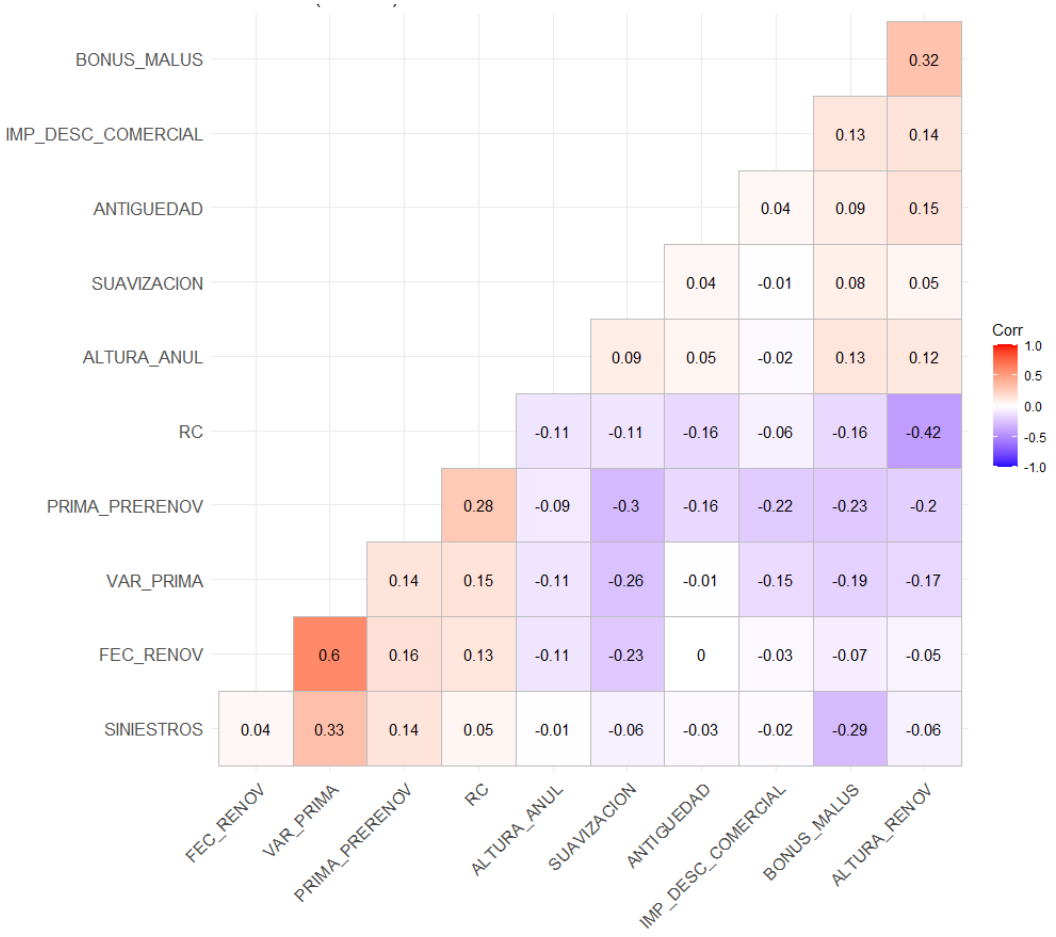

Tabla 8 Correlación Pearson (Numéricas) BBDD Sin correlaciones

Donde ya se observa que la data se encuentra sin correlaciones altas de las variables incluidas y analizadas en ella. De igual forma se retiran las variables de características como Fechas de Inicio y Fin, Importes de Primas, Datos de Identificación de los clientes y del vehículo.

Finalmente nos quedamos con 21 variables dentro de la base de datos, se muestra un análisis de las características de ellas:

```
TIPO_VEHICULO        MARCA            SINIESTROS        BONUS_MALUS       FEC_RENOV          llave            NEGOCIO            ANTIGUEDAD
Length:202238      Length:202238      Min.   :0.00000   Min.   :0.000   Min.   :202101   Length:202238      Length:202238      Min.   :  -1.000
Class :character   Class :character   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.:202108   Class :character   Class :character   1st Qu.:   2.000
Mode  :character   Mode  :character   Median :0.00000   Median :4.000   Median :202201   Mode  :character   Mode  :character   Median :   5.000
                                      Mean   :0.07928   Mean   :3.127   Mean   :202161                                          Mean   :   6.185
                                      3rd Qu.:0.00000   3rd Qu.:4.000   3rd Qu.:202207                                          3rd Qu.:   9.000
                                      Max.   :9.00000   Max.   :4.000   Max.   :202212                                          Max.   :2021.000
RANGO_ANTIGUEDAD   DEPARTAMENTO        MCA_RENOV            RC          REGIONAL         FIDELIZACION       ALTURA_RENOV     RANGO_PRIMA_ANT      VAR_PRIMA
Length:202238      Length:202238      Min.   :0.0000   Min.   : 110   Length:202238      Length:202238      Min.   : 1.000   Length:202238      Min.   :-1.00000
Class :character   Class :character   1st Qu.:1.0000   1st Qu.:2000   Class :character   Class :character   1st Qu.: 1.000   Class :character   1st Qu.:-0.01000
Mode  :character   Mode  :character   Median :1.0000   Median :3000   Mode  :character   Mode  :character   Median : 2.000   Mode  :character   Median : 0.08000
                                      Mean   :0.7841   Mean   :2296                                         Mean   : 2.611                      Mean   : 0.06891
                                      3rd Qu.:1.0000   3rd Qu.:3000                                         3rd Qu.: 3.000                      3rd Qu.: 0.14000
                                      Max.   :1.0000   Max.   :5000                                         Max.   :18.000                      Max.   : 2.41000
RANGO_SCORE          GENERO        RANGO_VAR_PRIMA
Length:202238      Length:202238      Length:202238
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

Tabla 9 Tabla de Características de las Variables Finales

5. RESULTADOS

GLM

Se realiza el modelo GLM inicial con todas las variables mencionadas en la Tabla 8, con el objetivo de extraer la información sobre que variables son o no significativas al modelo y pueden llegar a aportar más información a él y como se ha mencionado anteriormente siguiendo el principio fundamental de parsimonia.

```
glm(formula = MCA_RENOV ~ TIPO_VEHICULO + SINIESTROS + BONUS_MALUS +
    FEC_RENOV + NEGOCIO + MARCA + DEPARTAMENTO + RC + FIDELIZACION +
    ANTIGUEDAD + VAR_PRIMA + RANGO_ANTIGUEDAD + REGIONAL + ALTURA_RENOV +
    RANGO_PRIMA_ANT + RANGO_SCORE + GENERO + RANGO_VAR_PRIMA
    family = "binomial"
 data = Data_SinCorr)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.8285  0.3018  0.4763  0.6076  3.2030
```

Tabla 10 Resumen del Cálculo del GLM con Todas las Variables

Dado que la tabla es bastante extensa se adjuntan los resultados en el Excel haciendo referencia a la hoja *Tabla 10.*

A partir de este primer resumen del modelo, se muestra el coeficiente β para cada variable y su significancia. Recordemos que usualmente todo lo que tiene p-value menor a 0.001 es significante para los modelos, siendo estas:

- Tipo Vehículo
- Bonus Malus
- Fec Renov (Año-Mes)
- Fidelización
- Rango Antigüedad
- Regional
- Altura Renovación
- Rango Prima Anterior
- Genero

Se tiene en cuenta que para las variables categóricas que no se seleccionaron en el modelo cuando más de un nivel de alguna variable no arroja un p-value significativo para el modelo estas no las contemplamos en las variables del modelo seleccionado.

Se realiza la prueba nuevamente sobre el modelo con las variables anteriores, y se observa después de varios análisis que podemos quedarnos con las siguientes variables y mejoraremos las mediciones del modelo:

- Tipo Vehículo
- Bonus Malus
- Altura Renovación
- Rango Prima Anterior
- Genero

Obteniendo los siguientes resultados:

Call:
glm(formula = MCA_RENOV ~ TIPO_VEHICULO + BONUS_MALUS + ALTURA_RENOV +
RANGO_PRIMA_ANT + GENERO, family = "binomial", data = Data_SinCorr)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9071	0.3234	0.5113	0.5906	3.1504

Coefficients:

Estimate	Std.	Error	z value	Pr(>|z|)	
(Intercept)	0.974377	0.025064	38.876	<0.0000000000000002	***
TIPO_VEHICULO2	0.184186	0.014976	12.299	<0.0000000000000002	***
TIPO_VEHICULO3	0.020194	0.027469	0.735	0.46224	
BONUS_MALUS	0.200145	0.005379	37.211	<0.0000000000000002	***
ALTURA_RENOV	0.127746	0.003863	33.070	<0.0000000000000002	***
RANGO_PRIMA_ANT2.Entre1y1.5millones	-0.207625	0.018914	-10.977	<0.0000000000000002	***
RANGO_PRIMA_ANT3.Mayora1.5millones	-0.336352	0.021364	-15.744	<0.0000000000000002	***
GENEROIndefinido	-5.721.202	0.057262	-99.913	<0.0000000000000002	***
GENEROMasculino	-0.037278	0.013658	-2.729	0.00634	**
GENERONIT	0.335656	0.031622	10.614	<0.0000000000000002	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 211010 on 202237 degrees of freedom
Residual deviance: 155610 on 202228 degrees of freedom
AIC: 155630

Tabla 11 Modelo Logit con la selección de las variables

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	16.056	27.219
1	919	156.668

Accuracy : 0.8599
95% CI : (0.8584, 0.8614)
No Information Rate : 0.9155
P-Value [Acc > NIR] : 1

Kappa : 0.4684

Mcnemar's Test P-Value : <0.0000000000000002

Sensitivity : 0.94586
Specificity : 0.85198
Pos Pred Value : 0.37102
Neg Pred Value : 0.99417
Prevalence : 0.08451
Detection Rate : 0.07994
Detection Prevalence : 0.21545
Balanced Accuracy : 0.89892

'Positive' Class : 0

Tabla 12 Matriz de confusión y resultados para Tabla 10

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	15.601	28.065
1	327	158.245

Accuracy : 0.8596
95% CI : (0.8581, 0.8611)
No Information Rate : 0.9212
P-Value [Acc > NIR] : 1

Kappa : 0.4614

Mcnemar's Test P-Value : <0.0000000000000002

Sensitivity : 0.97947
Specificity : 0.84936
Pos Pred Value : 0.35728
Neg Pred Value : 0.99794
Prevalence : 0.07876
Detection Rate : 0.07714
Detection Prevalence : 0.21591
Balanced Accuracy : 0.91442

'Positive' Class : 0

Tabla 13 Matriz de Confusión y resultados para Tabla 11

Una vez validado el modelo a partir de toda la base de Datos se realiza la partición de la data 70% - 30% entre train y test, adicionalmente para la base de test se corta la información a los últimos 5 meses del año 2022, esto realiza con el fin de validar si las predicciones hechas anteriormente son válidas también para los últimos periodos del año, en los cuales la situación de los rezagos de la pandemia de Covid-19 ya no son tan fuertes en el país y la variación de las primas no es tan volátil como en meses anteriores. A continuación, se muestran los resultados del proceso:

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	10.838	19.632
1	221	110.875

Accuracy : 0.8598
95% CI : (0.8579, 0.8616)
No Information Rate : 0.9219
P-Value [Acc > NIR] : 1

Kappa : 0.4601

Mcnemar's Test P-Value :
<0.0000000000000002

Sensitivity : 0.98002
Specificity : 0.84957
Pos Pred Value : 0.35569
Neg Pred Value : 0.99801
Prevalence : 0.07812
Detection Rate : 0.07656
Detection Prevalence : 0.21524
Balanced Accuracy : 0.91479

'Positive' Class : 0

Tabla 14 Resultados del modelo en la base de Train

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	906	1.820
1	33	11.983

Accuracy : 0.8743
95% CI : (0.8688, 0.8796)
No Information Rate : 0.9363
P-Value [Acc > NIR] : 1

Kappa : 0.4415

Mcnemar's Test P-Value :
<0.0000000000000002

Sensitivity : 0.96486
Specificity : 0.86814
Pos Pred Value : 0.33236
Neg Pred Value : 0.99725
Prevalence : 0.06370
Detection Rate : 0.06146
Detection Prevalence : 0.18491
Balanced Accuracy : 0.91650

'Positive' Class : 0

Tabla 15 Resultados del modelo en la base de Train para >= 202208

Una vez validado que el modelo tiene factor de exactitud alto y los demás factores son significativos para el modelo, procedemos a realizar el *Cross-validation* desde la partición de las bases de train y test. Obteniendo así los siguientes resultados de esta, se reduce la muestra a los últimos 3 meses de información:

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	1.795	3.943
1	45	25.930

Accuracy : 0.8742
95% CI : (0.8705, 0.8779)
No Information Rate : 0.942
P-Value [Acc > NIR] : 1

Kappa : 0.423

Mcnemar's Test P-Value :
<0.0000000000000002

Sensitivity : 0.97554
Specificity : 0.86801
Pos Pred Value : 0.31283
Neg Pred Value : 0.99827
Prevalence : 0.05802
Detection Rate : 0.05660
Detection Prevalence : 0.18094
Balanced Accuracy : 0.92178

'Positive' Class : 0

Tabla 16 Resultados del Modelo GLM de Cross Validation

De acuerdo con los resultados del modelo final sobre la data de test del cross validation, se procede a ordenar la data de menor a mayor probabilidad predicha por el modelo y se parte la base de datos en 20 grupos (cada uno con el 5% de los datos) con el fin de poder observar la tendencia de la probabilidad predicha por el modelo:

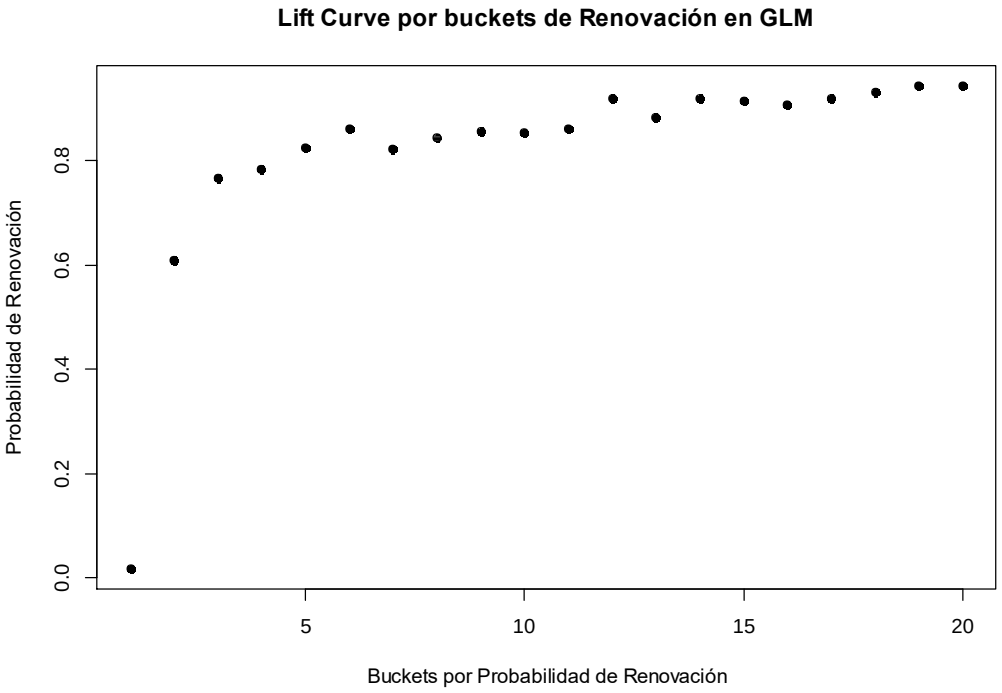

Ilustración 41 Lift Chart Modelo GLM

Sobre la cual se observa que la probabilidad predicha por el modelo genera buckets de la cartera los cuales tienen una tendencia de probabilidad de renovación creciente de acuerdo con la combinación de todas sus variables. De igual manera, se tiene un 5% con probabilidad de renovación cercana al 0% este segmento esta principalmente relacionado con la variable género, allí se encuentra bastante información sobre el género *indefinido,* sobre los cuáles a partir de juicio de experto y conocimiento en la compañía podríamos determinar que durante los procesos de renovación pueden generar incumplimiento de los requisitos de suscripción de la compañía, por lo cual en su mayoría no renuevan por falta de información veraz para la compañía.

A continuación, se generará la validación de la probabilidad predicha y la probabilidad real para cada una de las variables de la base de datos (tanto las variables usadas en el GLM final como el restante de variables relevantes para el riesgo).

NOTA: Al final de los gráficos se genera el análisis resumen sobre lo observado en ellos.

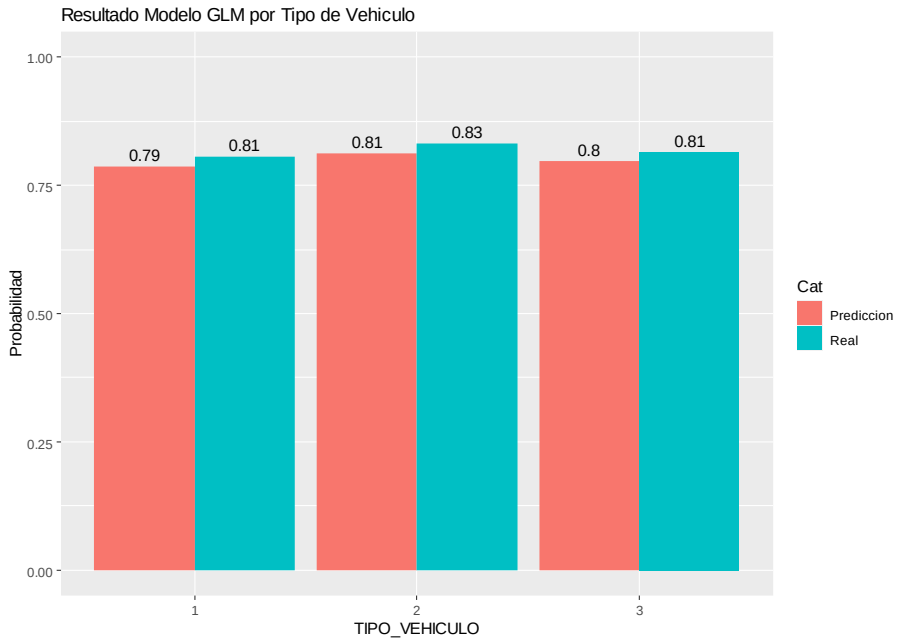

Ilustración 42 Resultado GLM por Tipo de Vehículo

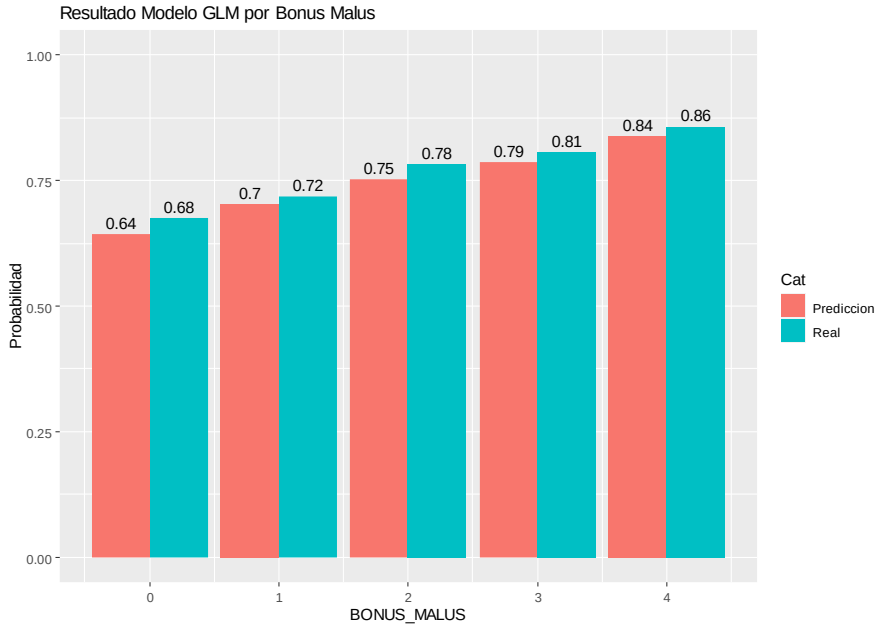

Ilustración 43 Resultado GLM por Bonus Malus

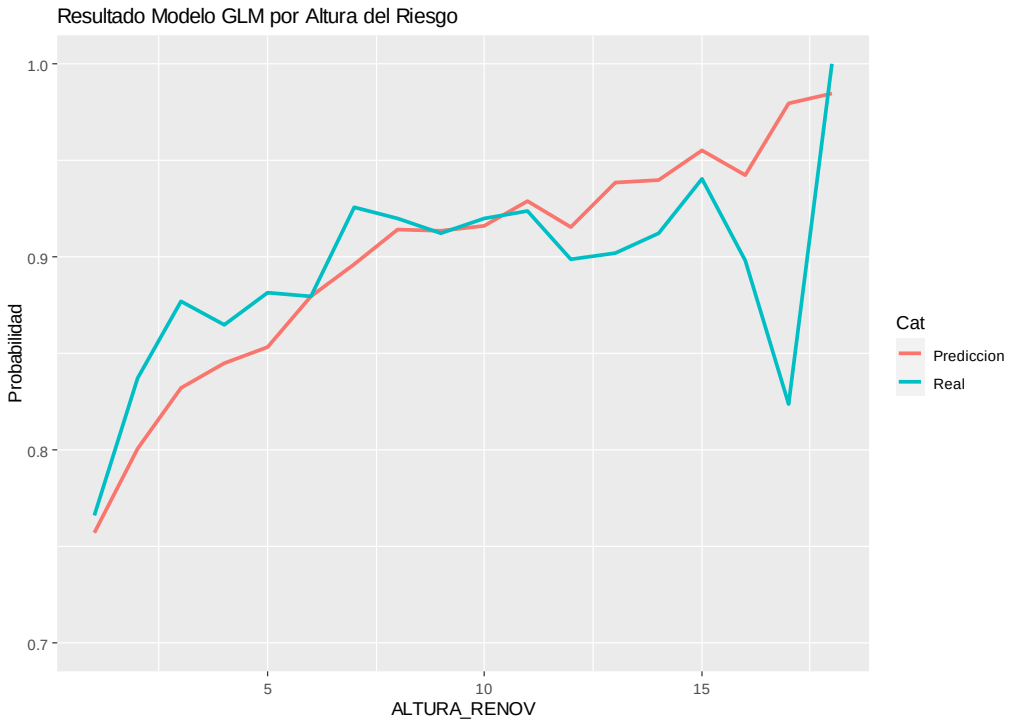

Ilustración 44 Resultado GLM por Altura de Renovación

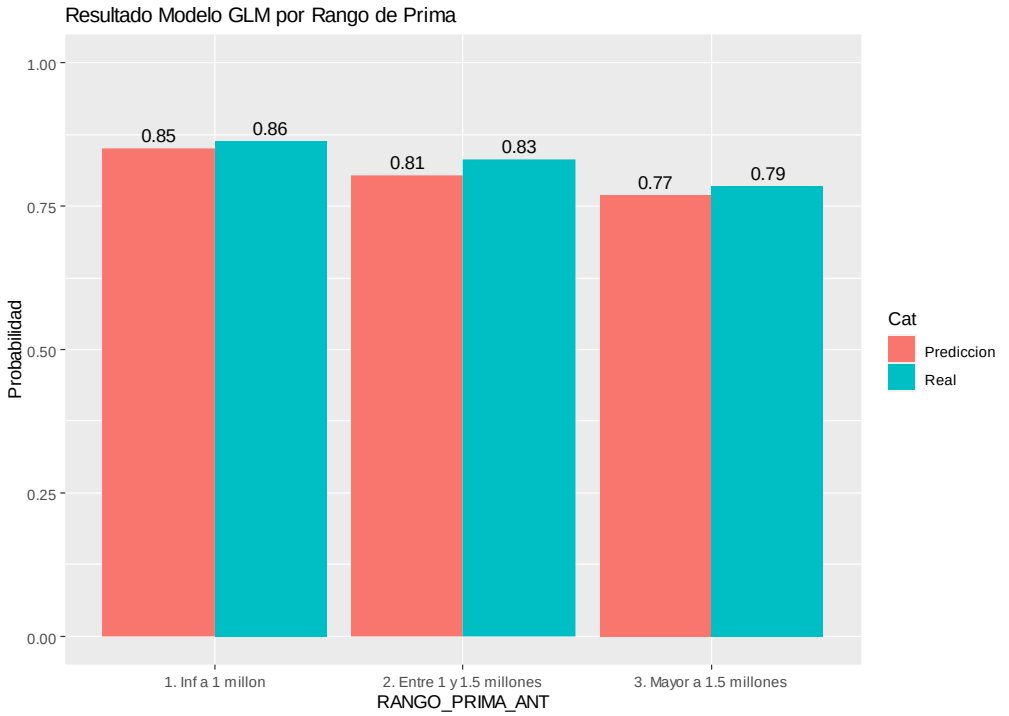

Ilustración 45 Resultado GLM por Rango de Prima Anterior

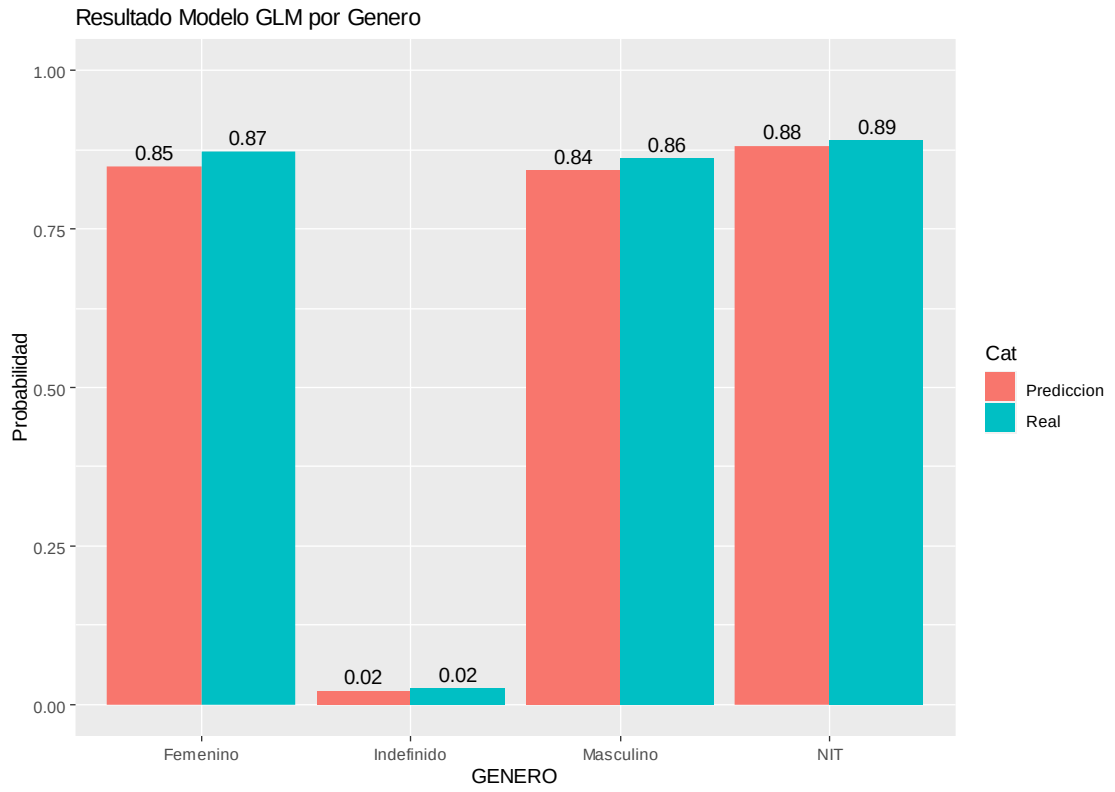

Ilustración 46 Resultado GLM por Genero

Ilustración 47 Resultado GLM por Fecha de Renovación

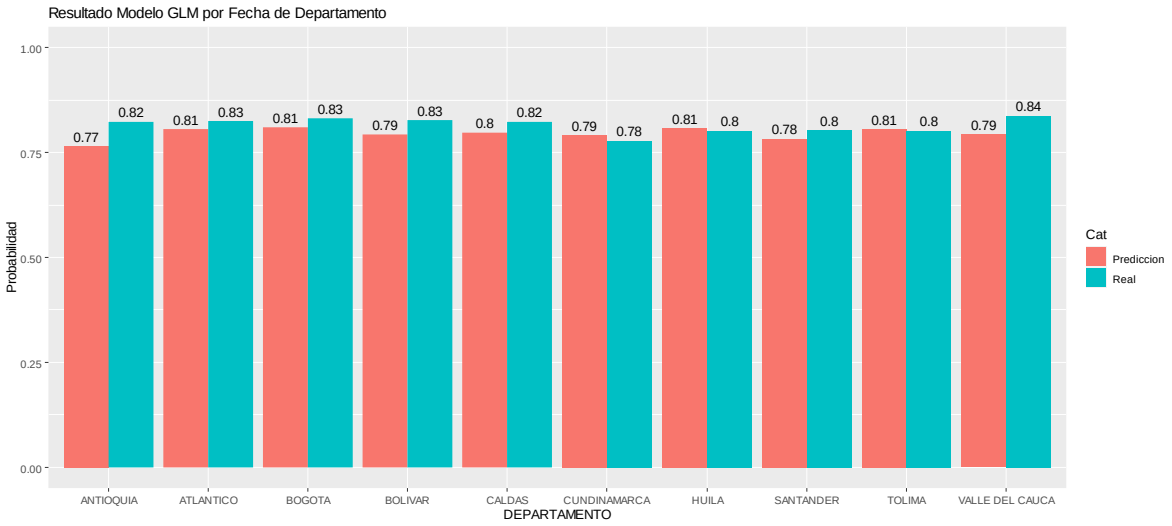

Ilustración 48 Resultado GLM por Departamento

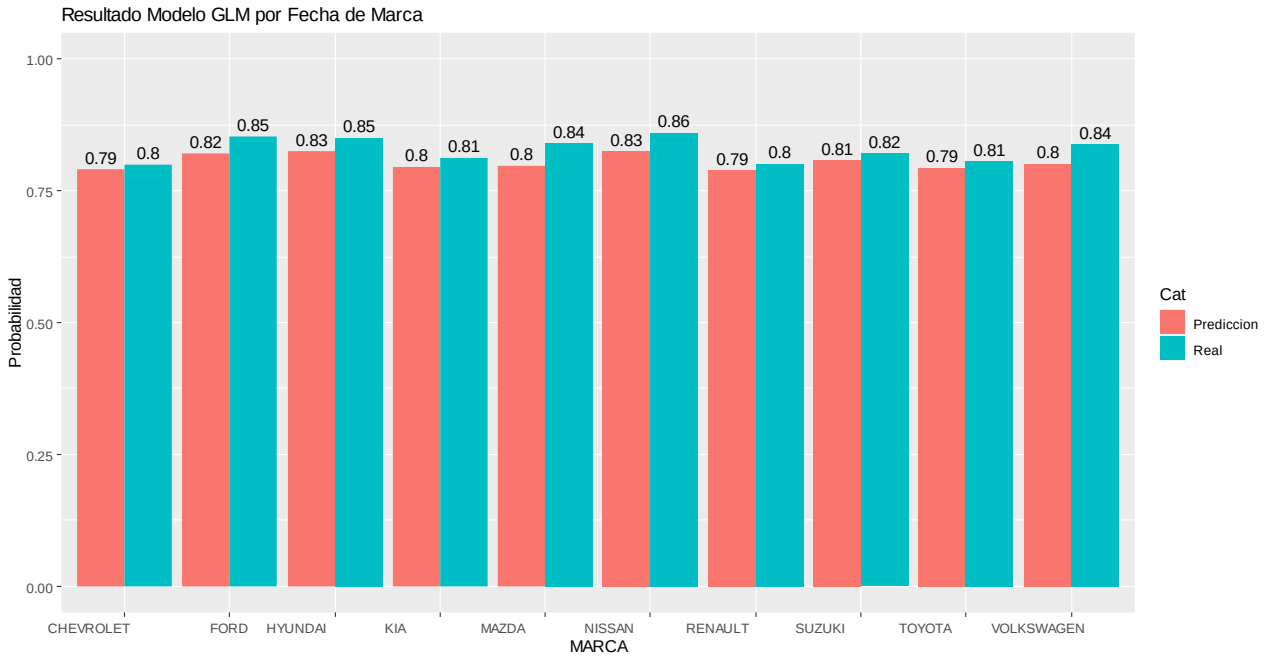

Ilustración 49 Resultado GLM por Marca

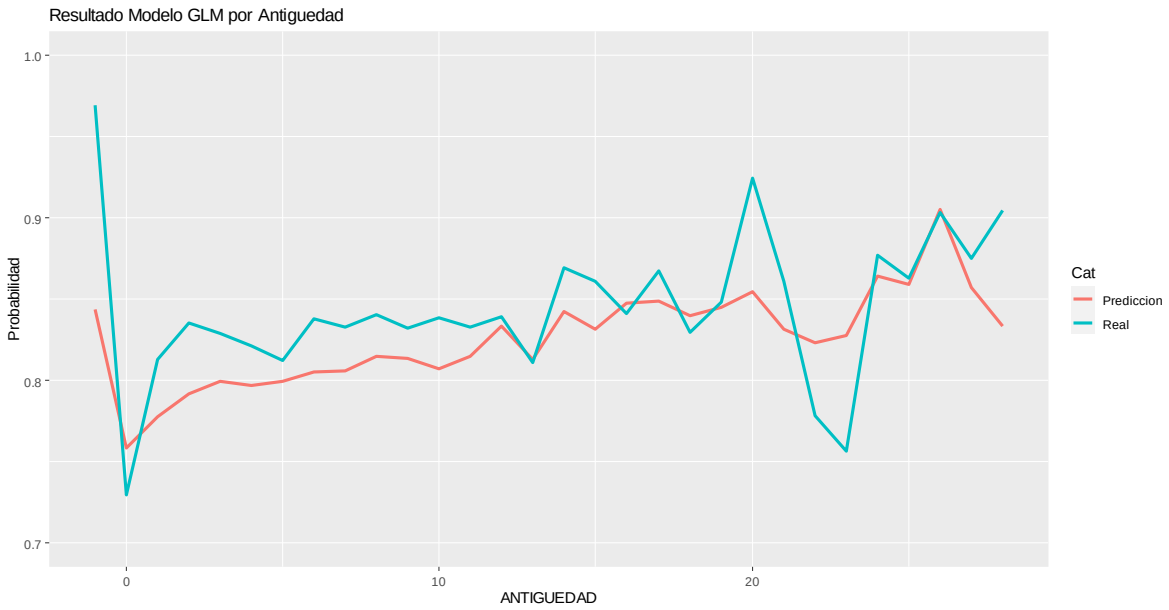

Ilustración 50 Resultado GLM por Antigüedad

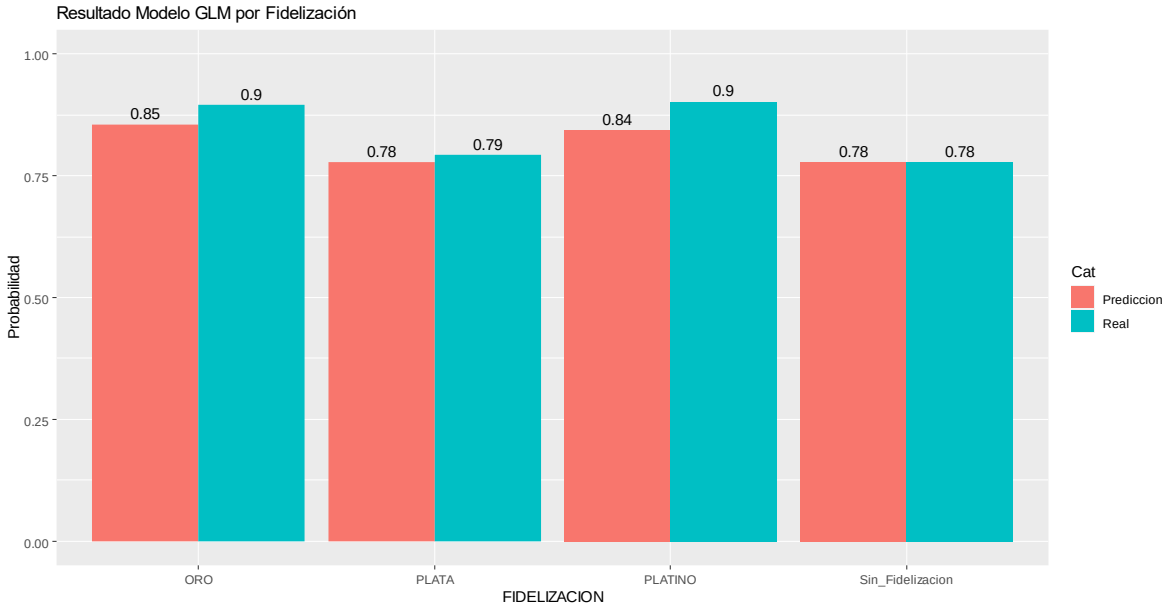

Ilustración 51 Resultado GLM por Fidelización

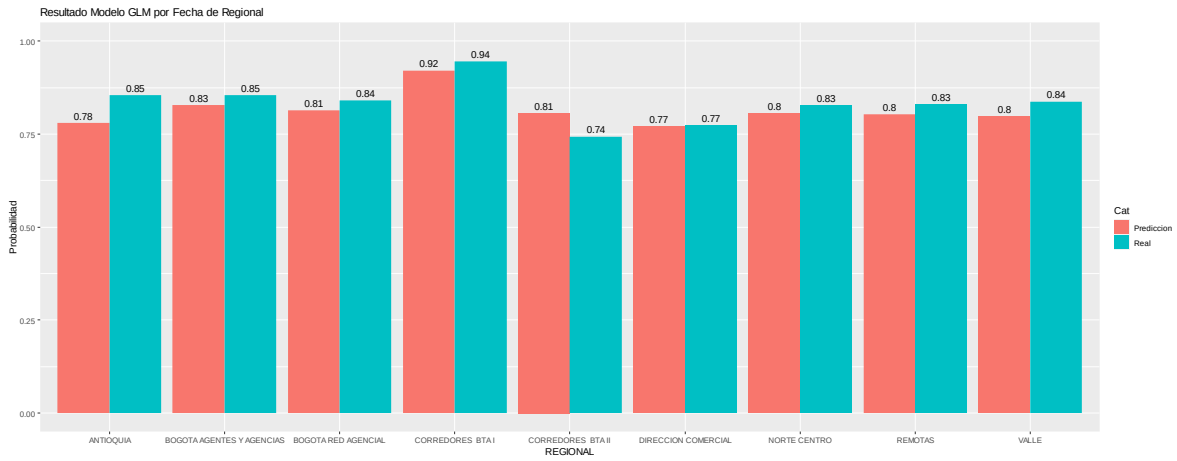

Ilustración 52 Resultado GLM por Regional

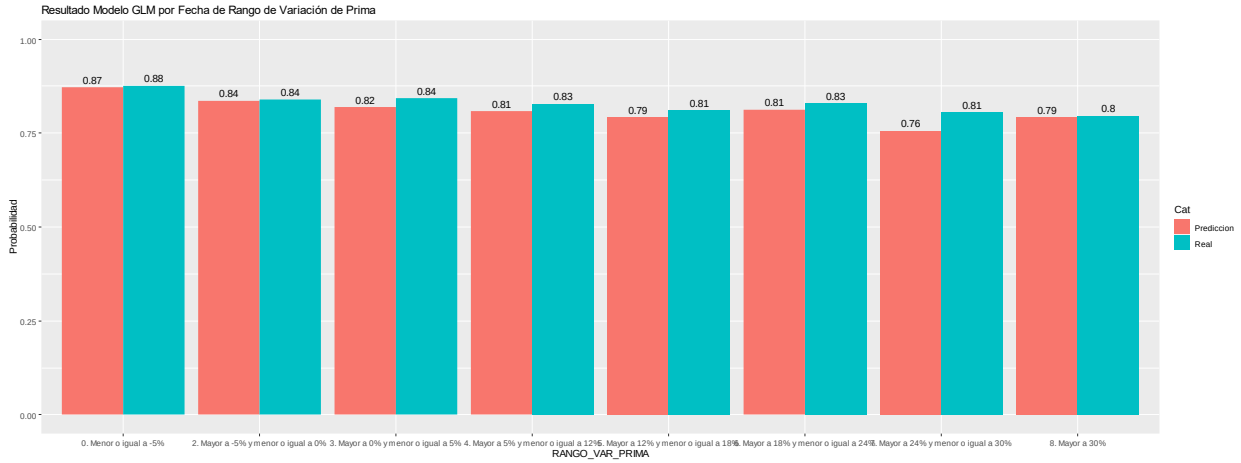

Ilustración 53 Resultado GLM por Variación de Prima

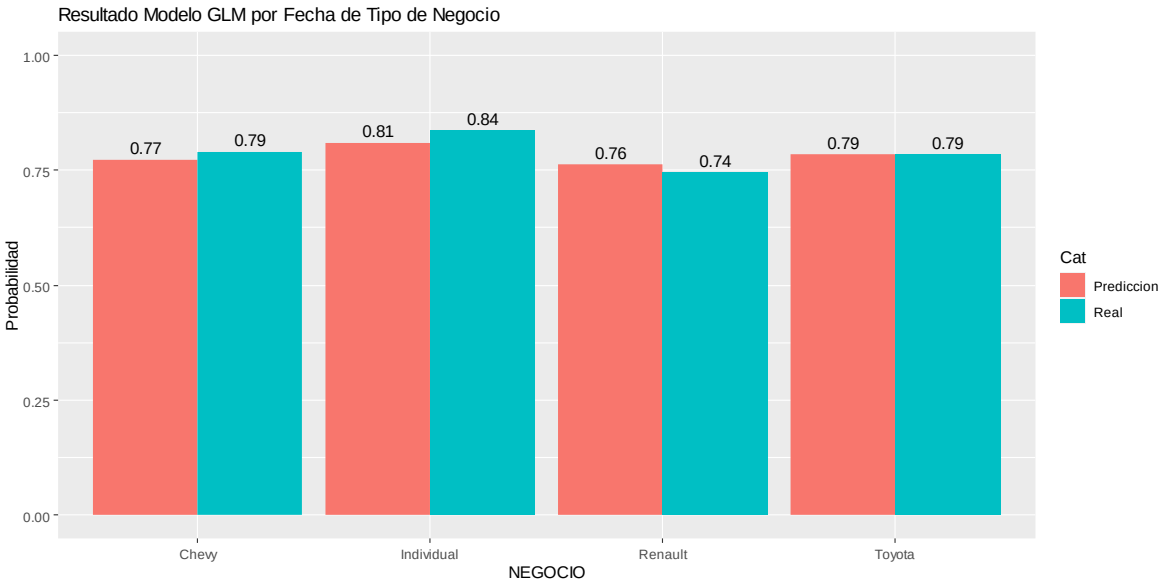

Ilustración 54 Resultado GLM por Negocio

Observando cada una de las predicciones y los valores reales de la probabilidad de renovación para cada una de las variables de la base de datos, podemos darnos cuenta de que el modelo GLM se ajusta bastante bien a la probabilidad de renovación por segmento, este mantiene una probabilidad por debajo o muy cercana a la real, lo cual nos permite hacer estimaciones con mayor seguridad sobre aquellos riesgos de probabilidad baja, pero de interés para la compañía. De acuerdo con este análisis, es posible afirmar que el modelo genera ajustes bastante buenos en sus predicciones como se puede observar por ejemplo en la *ilustración 46* de la altura de renovación, el gráfico de predicción sigue la tendencia que se emplea normalmente al momento de generar modelos de tarifa, ajustándose a los valores reales y suavizando la curva en categorías que generar saltos en la información. En el caso de la *ilustración 46* por género, es claro que el género *Indefinido* está determinando con bastante fuerza aquellos riesgos con probabilidades casi nulas de no renovación, sobre esta variable se tomarán decisiones dentro de la compañía con el fin de no generar un poco de sesgo las predicciones del modelo, pues ya que no son datos que se puedan retirar directamente de la base, esto generaría un incremento significativo sobre los indicadores de retención lo cuál sería un dato errado frente a los datos de la compañía.

Random Forest

Para la ejecución del Random Forest, se ejecuta inicialmente todo el modelo con todas las variables con el fin de poder observar la importancia de estas variables para el Random forest, se obtienen los siguientes resultados:

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	40.895	1.293
1	2.771	157.279

Accuracy : 0.9799
95% CI : (0.9793, 0.9805)
No Information Rate : 0.7841
P-Value [Acc > NIR] : < 0.00000000000000022

Kappa : 0.9399

Mcnemar's Test P-Value : <
0.00000000000000022

Sensitivity : 0.9365
Specificity : 0.9918
Pos Pred Value : 0.9694
Neg Pred Value : 0.9827
Prevalence : 0.2159
Detection Rate : 0.2022
Detection Prevalence : 0.2086
Balanced Accuracy : 0.9642

'Positive' Class : 0

Tabla 17 Resultado del Random Forest con Todas las Variables

Se observa la significancia del modelo es bastante alta al comparar con el modelo del GLM, sin embargo, hay que tener presente que los random forest a veces cuando se hacen con las bases iniciales y de train suelen dar muy buenos resultados y a veces con el test la significancia del modelo disminuye considerablemente. Adicionalmente se observan las variables cuya importancia para el random forest son clave en el modelo, una de ella principalmente es el género, la fecha de renovación y la variación de prima:

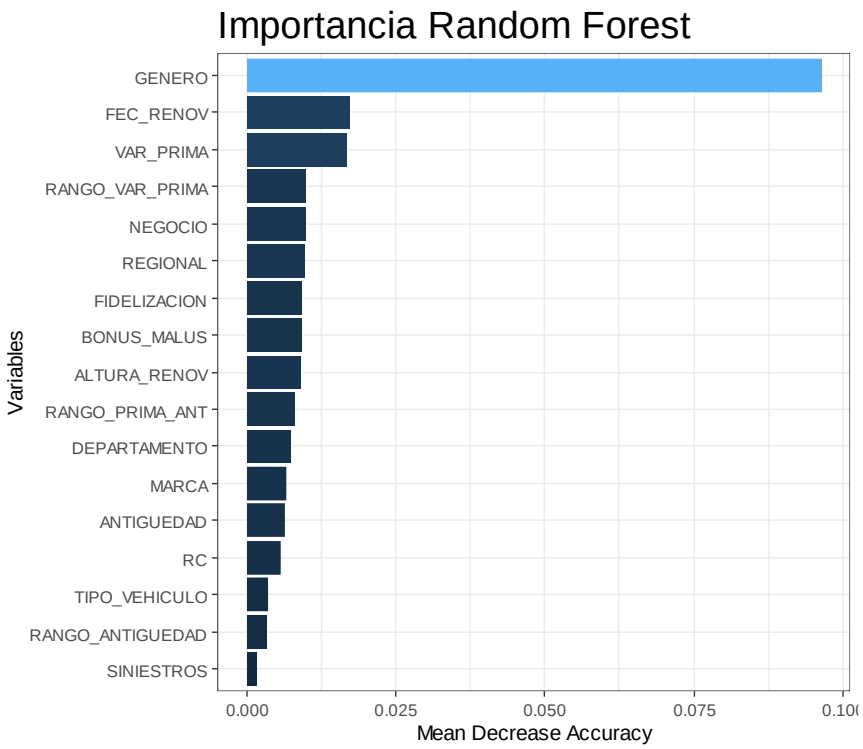

Tabla 18 Importancia de las variables en el Random Forest

Dado que la variación de prima y el rango de variación de prima están correlacionadas, seleccionaremos la variación de prima, de igual forma con la antigüedad y el rango de antigüedad. Al seleccionar las siguientes variables y ajustar los hiperparametros del modelo se obtienen los resultados:

- Bonus Maus
- Fecha de Renovación
- Negocio
- Marca
- Departamento
- Fidelización
- Antigüedad
- Variación Prima
- Regional
- Altura renovación
- Rango Prima Anterior
- Genero

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	37.954	3.495
1	5.712	155.077

Accuracy : 0.9545
95% CI : (0.9536, 0.9554)
No Information Rate : 0.7841
P-Value [Acc > NIR] : < 0.00000000000000022

Kappa : 0.863

Mcnemar's Test P-Value : < 0.00000000000000022

Sensitivity : 0.8692 Specificity : 0.9780
Pos Pred Value : 0.9157
Neg Pred Value : 0.9645
Prevalence : 0.2159
Detection Rate : 0.1877
Detection Prevalence : 0.2050
Balanced Accuracy : 0.9236

'Positive' Class : 0

Tabla 19 Resultados Random Forest seleccionando variables

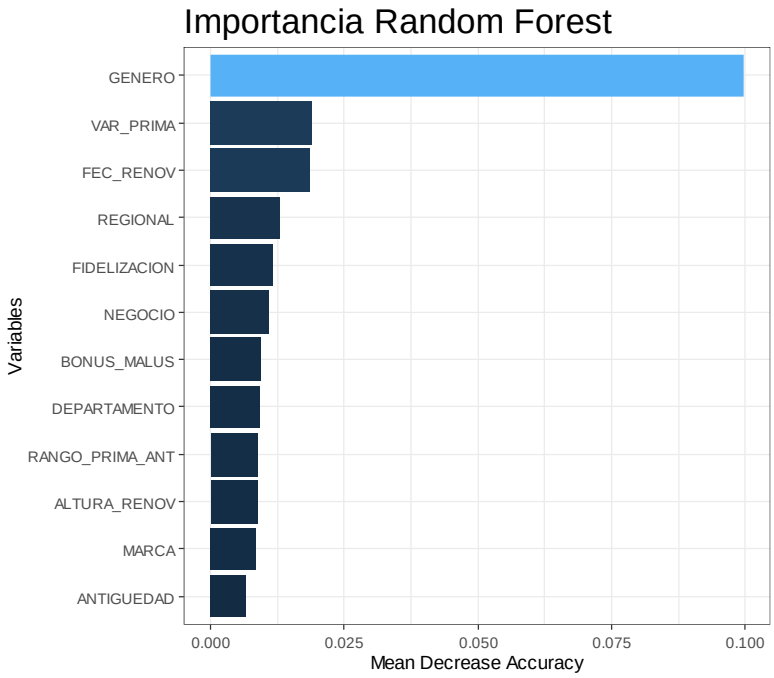

Tabla 20 Importancia Random Forest con selección de variables

Una vez seleccionados los hiperparametros y las variables a usar en el modelo de random forest, realizamos el cross validation tanto para la partición de la base en train y test como en la construcción de los árboles. De la misma manera que cortamos el testo para el modelo GLM usando los últimos 3 meses de información con el fin de observar la medición del modelo en un escenario más neutro. Se obtienen los siguientes resultados.

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	2.014	58
1	3.724	25.917

Accuracy : 0.8807
95% CI : (0.8771, 0.8843)
No Information Rate : 0.8191
P-Value [Acc > NIR] : <
0.00000000000000022

Kappa : 0.4643

Mcnemar's Test P-Value : <
0.00000000000000022

Sensitivity : 0.35099
Specificity : 0.99777
Pos Pred Value : 0.97201
Neg Pred Value : 0.87436
Prevalence : 0.18094
Detection Rate : 0.06351
Detection Prevalence : 0.06534
Balanced Accuracy : 0.67438

'Positive' Class : 0

Tabla 21 Resultados del Modelo de Random Forest con Cross Validation

Una vez generado el cross validation sobre la base de test procedemos al realizar el mismo análisis que hicimos para el modelo GLM, se procede a ordenar la data de menor a mayor probabilidad predicha por el modelo y se parte la base de datos en 20 grupos (cada uno con el 5% de los datos) con el fin de poder observar la tendencia de la probabilidad predicha por el modelo:

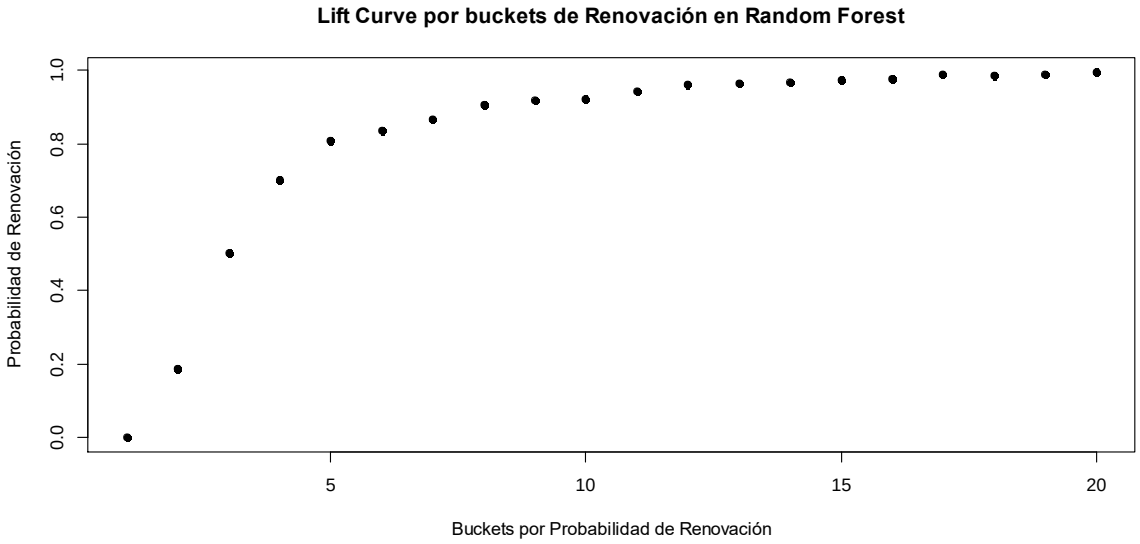

Ilustración 55 Lift Chart Modelo Random Forest

Sobre la cual se observa que la probabilidad predicha por el modelo genera buckets de la cartera los cuales tienen una tendencia de probabilidad de renovación creciente de acuerdo con la combinación de todas sus variables. De igual manera, se tiene un 5% con probabilidad de renovación cercana al 0% este segmento esta principalmente relacionado con la variable género, y otro 5% con probabilidades cercanas al 20% allí se encuentra bastante información sobre el género *indefinido,* sobre los cuáles a partir de juicio de experto y conocimiento en la compañía podríamos determinar que durante los procesos de renovación pueden generar incumplimiento de los requisitos de suscripción de la compañía, por lo cual en su mayoría no renuevan por falta de información veraz para la compañía. Sin embargo, a comparación del modelo GLM el random forest, suaviza un poco más la curva de probabilidades de renovación generando más buckets con probabilidades menores al 70% de renovación. Y en los últimos grupos probabilidades casi que del 100% de probabilidad de renovación (el bucket con mayor probabilidad en el GLM era del 95%).

A continuación, se generará la validación de la probabilidad predicha y la probabilidad real para cada una de las variables de la base de datos (tanto las variables usadas en el random forest final como el restante de variables relevantes para el riesgo).

NOTA: Al final de los gráficos se genera el análisis resumen sobre lo observado en ellos.

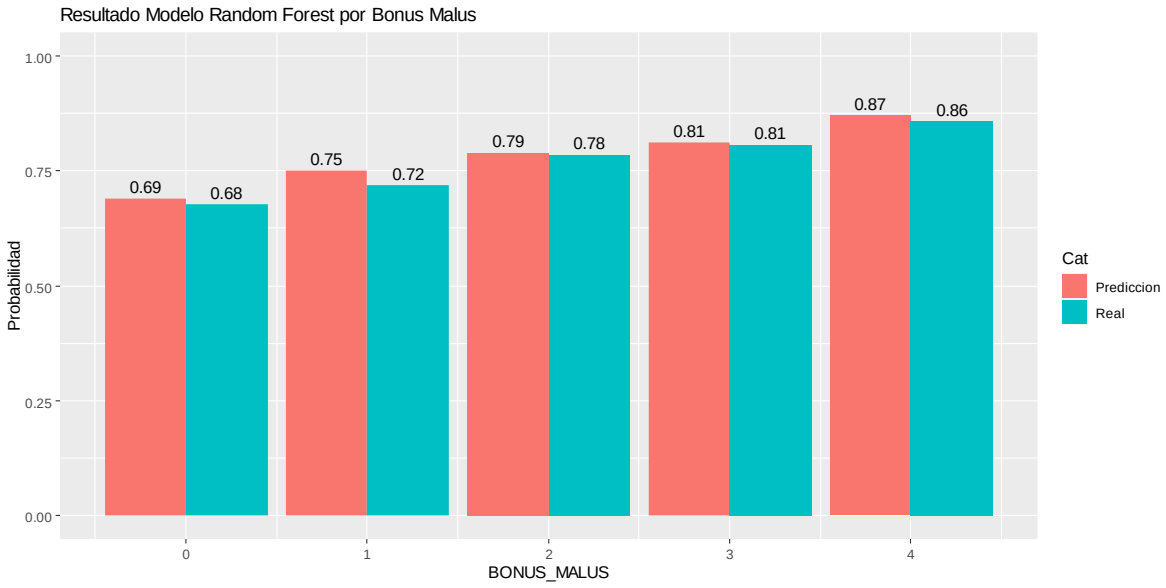

Ilustración 56 Resultado Random Forest Por Bonus Malus

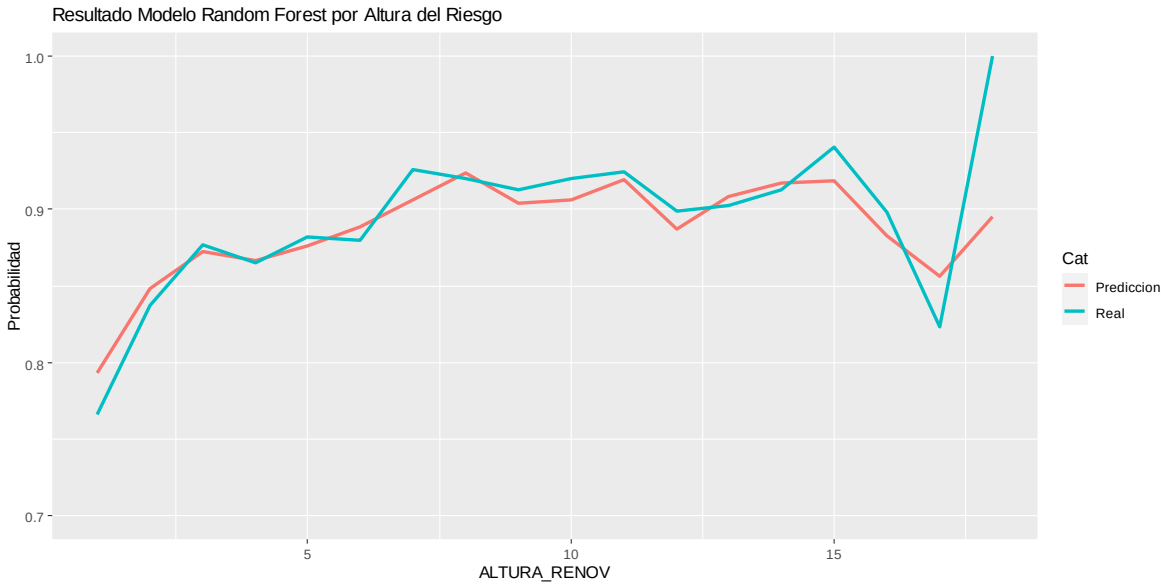

Ilustración 57 Resultado Random Forest Por Altura de Renovación

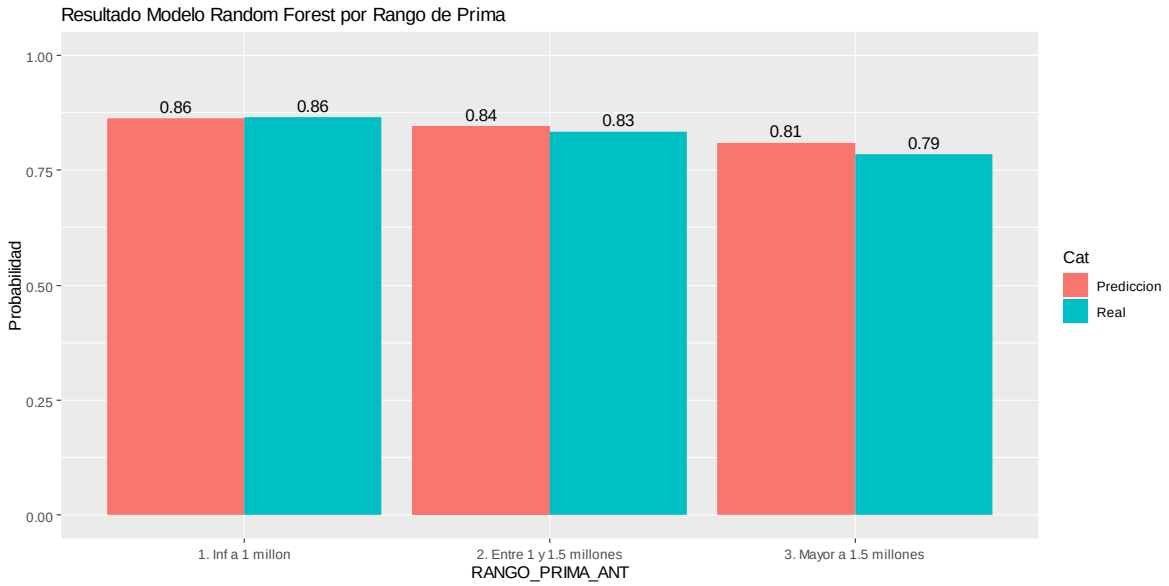

Ilustración 58 Resultado Random Forest Por Rango de Prima Anterior

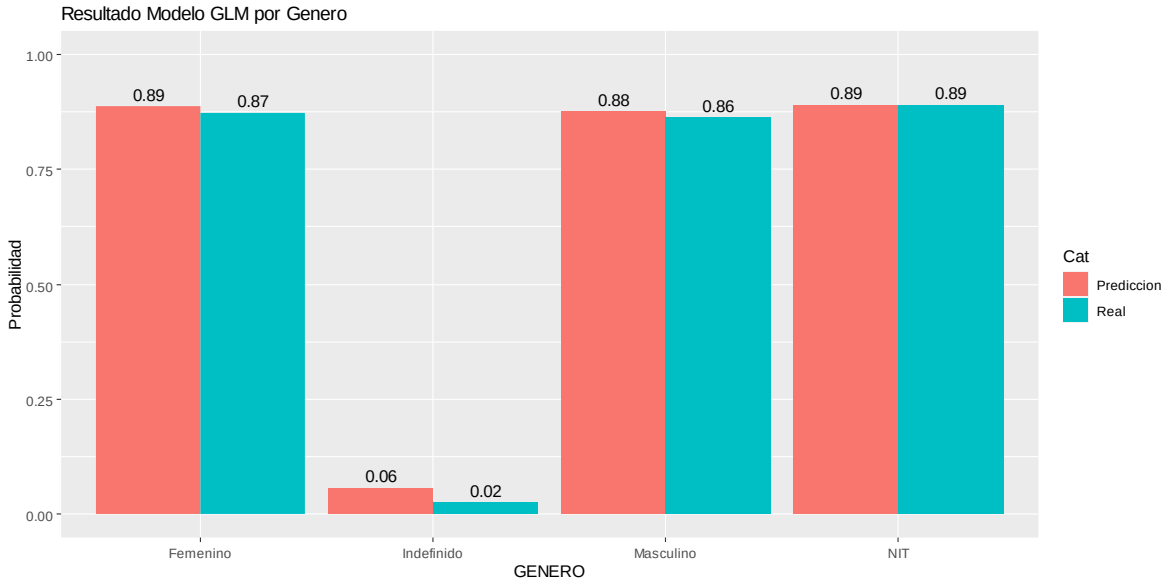

Ilustración 59 Resultado Random Forest Por Genero

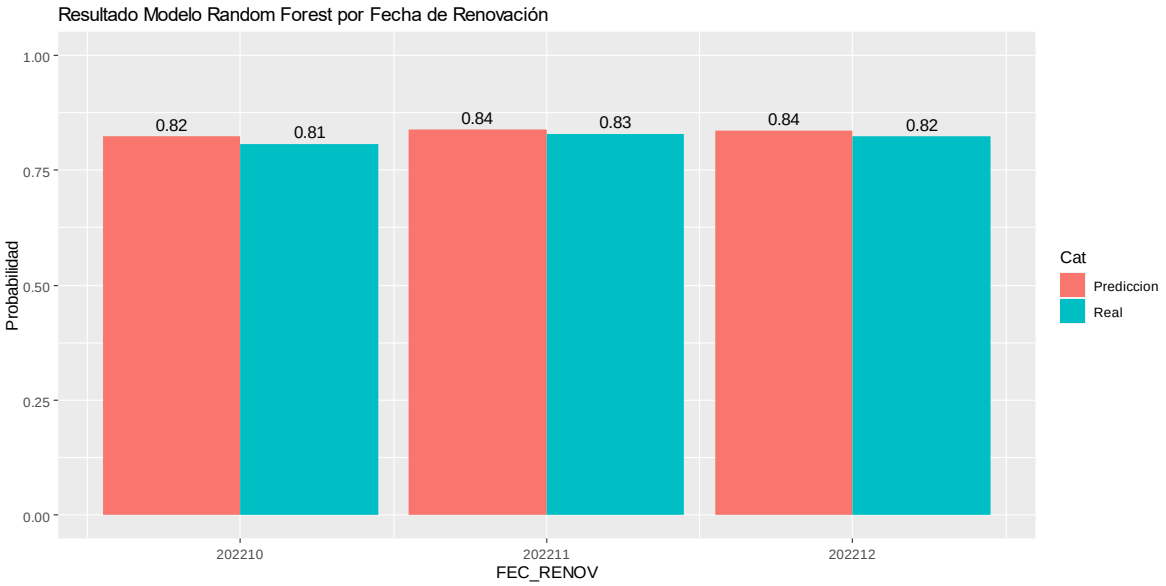

Ilustración 60 Resultado Random Forest Por Fecha de Renovación

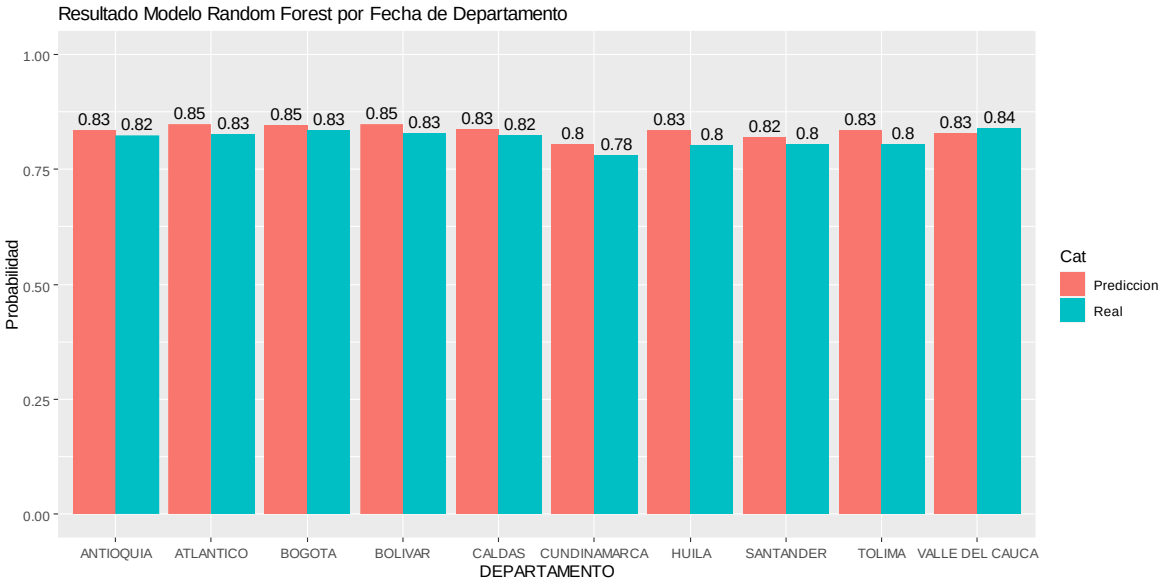

Ilustración 61 Resultado Random Forest Por Departamento

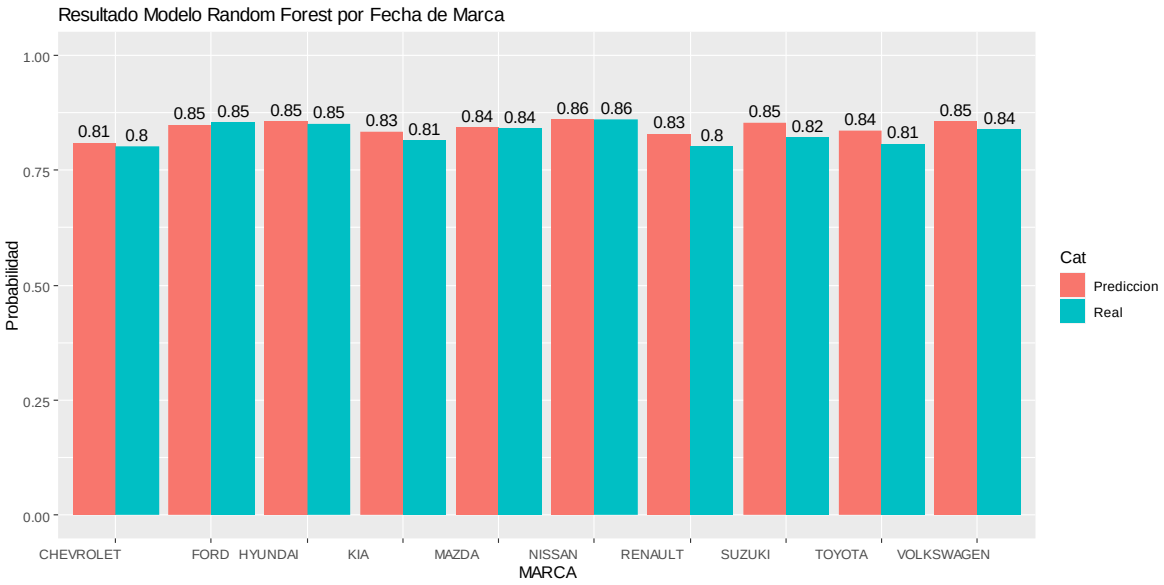

Ilustración 62 Resultado Random Forest Por Marca

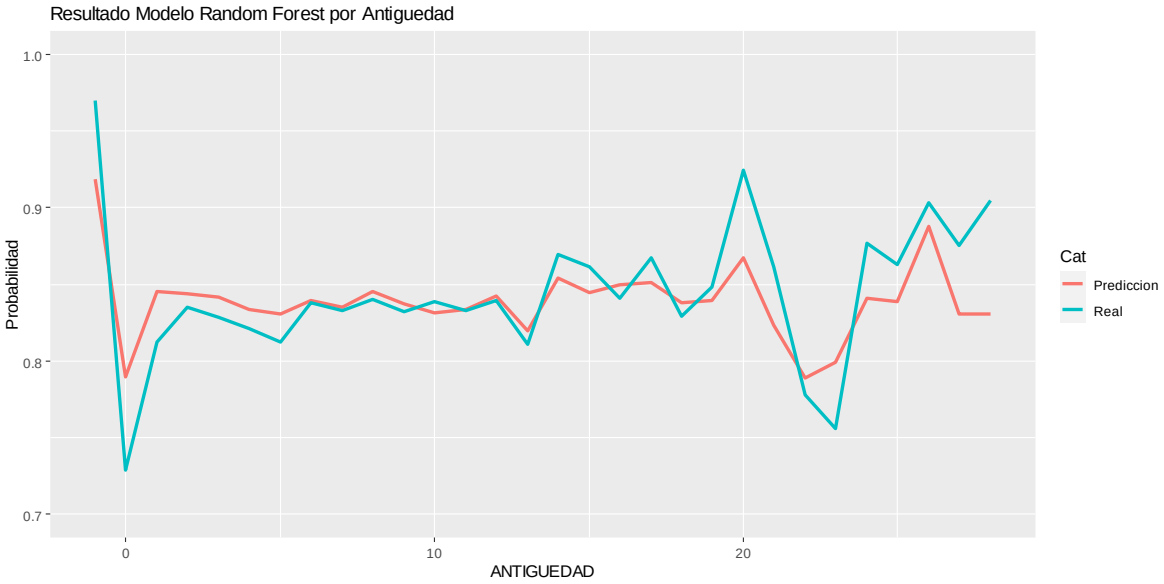

Ilustración 63 Resultado Random Forest Por Antigüedad

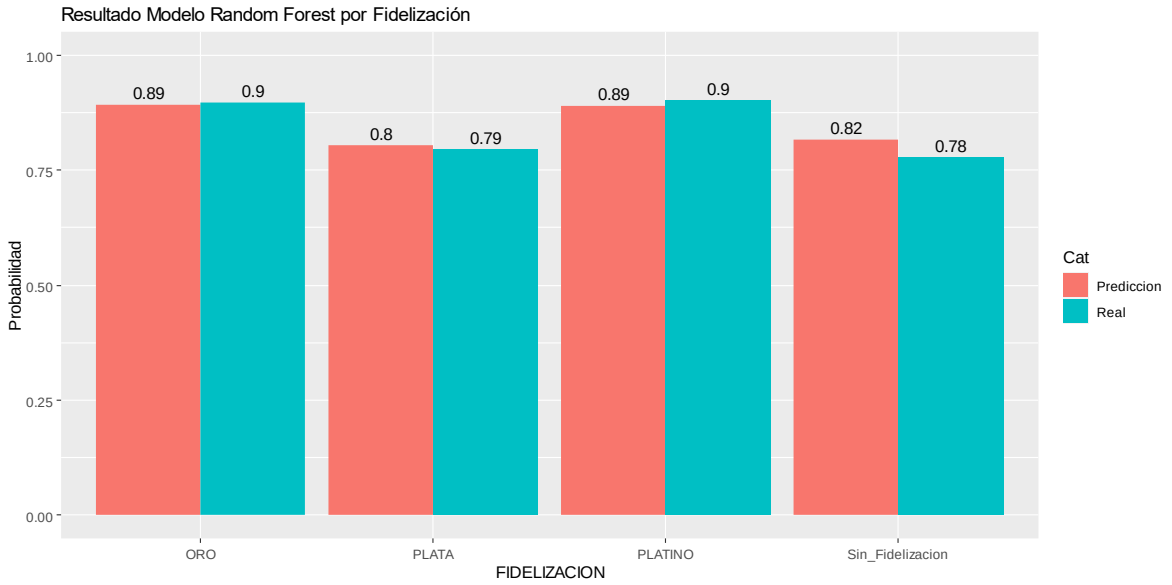

Ilustración 64 Resultado Random Forest Por Fidelización

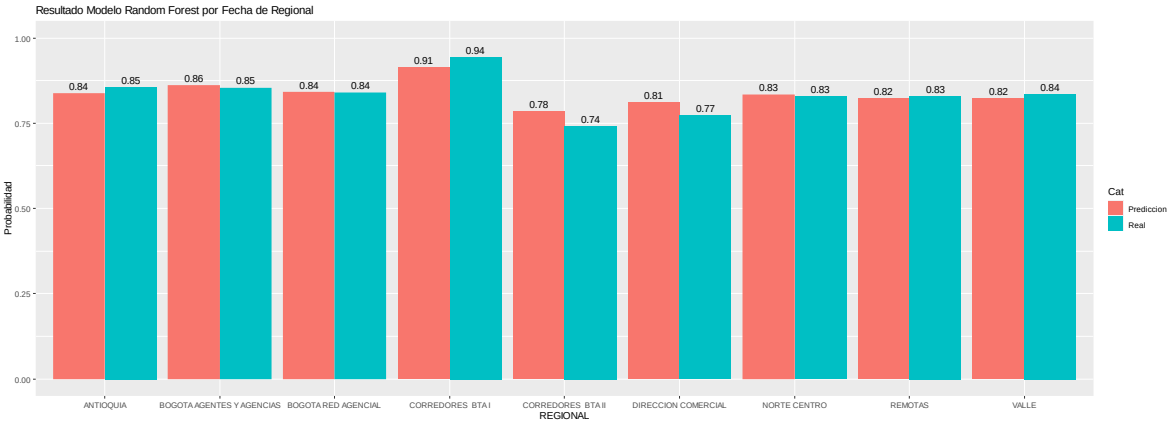

Ilustración 65 Resultado Random Forest Por Regional

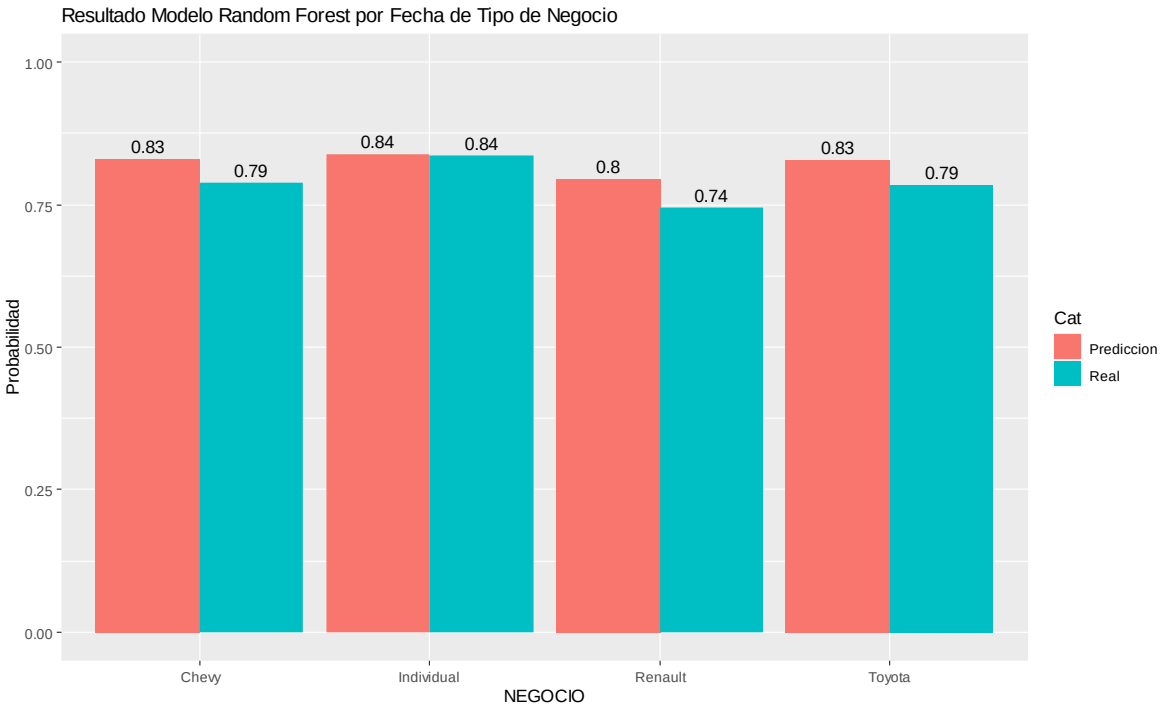

Ilustración 66 Resultado Random Forest Por Negocio

Observando cada una de las predicciones y los valores reales de la probabilidad de renovación para cada una de las variables de la base de datos, podemos darnos cuenta de que el modelo Random Forest genera un ligero sobreajuste a la probabilidad de renovación frente al dato real. Lo cual nos puede sesgar un poco las estimaciones de aquellos riegos con probabilidades bajas de renovación, pues si bien el dato real de renovación se encuentra cercano al 79% en la predicción del random forest llega a superar el 83%. Esto se muestra en el bajo indicie de sensibilidad que genera el modelo, de esta manera a juicio se considera que el modelo a pesar de generar estimaciones racionales contra el dato real, este no es el mejor modelo por elegir para un proceso de optimización de la compañía.

6. CONCLUSIONES

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	1.795	3.943
1	45	25.930

Accuracy : 0.8742
95% CI : (0.8705, 0.8779)
No Information Rate : 0.942
P-Value [Acc > NIR] : 1

Kappa : 0.423

Mcnemar's Test P-Value :
<0.0000000000000002

Sensitivity : 0.97554
Specificity : 0.86801
Pos Pred Value : 0.31283
Neg Pred Value : 0.99827
Prevalence : 0.05802
Detection Rate : 0.05660
Detection Prevalence : 0.18094
Balanced Accuracy : 0.92178

'Positive' Class : 0

Tabla 23 GLM con Cross Validation

Confusion Matrix and Statistics

	Reference	
Prediction	**0**	**1**
0	2.014	58
1	3.724	25.917

Accuracy : 0.8807
95% CI : (0.8771, 0.8843)
No Information Rate : 0.8191
P-Value [Acc > NIR] : <
0.00000000000000022

Kappa : 0.4643

Mcnemar's Test P-Value : <
0.00000000000000022

Sensitivity : 0.35099
Specificity : 0.99777
Pos Pred Value : 0.97201
Neg Pred Value : 0.87436
Prevalence : 0.18094
Detection Rate : 0.06351
Detection Prevalence : 0.06534
Balanced Accuracy : 0.67438

'Positive' Class : 0

Tabla 22 Random Forest con Cross Validation

Al realizar un comparativo entre las tablas 22 y 23 respectivamente, las cuales son los resultados finales de cada uno de los modelos, se puede observar que a pesar de que la exactitud del modelo (accuracy) y el Kappa para el Random Forest son ligeramente más altos que el GLM, observamos los niveles de sensibilidad y especificidad para el modelo de GLM tienen un mejor resultado. Así mismo es necesario tener en cuenta que el modelo GLM tiene m menos variables dentro del modelo con el fin de seguir el principio de parsimonia dentro de este. Al mismo nivel, se considera de acuerdo con los análisis de validación que la estimación de la probabilidad de renovación genera una mejor estimación el modelo GLM que el Random Forest para la compañía.

Por lo cual podemos decir que tenemos dos buenos modelos capaces de predecir de un cierto modo la probabilidad de renovación de los riesgos asegurados para el ramo de automóviles. De esta forma podemos llegar a medir de una forma más optima la

caída de cartera del negocio, de tal manera que nos enfocaremos en estos riesgos con probabilidades de no renovar pero que son de gran interés para la compañía. Si son modelos que a medida que se pueden ir robusteciendo con más variables pueden llegar a ser mucho más exactos en los análisis y aportar mayor valor a la compañía. Además si llegásemos a poder recolectar variables exógenas a la tarifa del cliente, tales como KPI's de servicio al cliente en el momento de una asistencia y/o un accidente, cantidad de productos dentro de la compañía en ramos distintos a automóviles, multas o infracciones durante el último año vigentes, etc., podemos llegar a robustecer mucho más el modelo y llegar a optimizar de forma correcta los precios, esto junto con modelos de elasticidad de precios para el mercado pueden llegar a convertirse en los pilotos de optimización para algunas compañías aseguradoras en Colombia.

Ya que el mercado asegurador de autos en Colombia ha logrado estabilizarse nuevamente después de la pandemia de COVID-19, se tiene en claro que el año 2024 es un año clave para las compañías, ya que el reto después de los fuertes incrementos en tarifa será lograr mantener la cartera asegurada con unos buenos niveles de renovación y generando rentabilidad para la empresa, la clave está en lograr retener buenos clientes que lleguen a tener probabilidades de caída alta y me aporten solvencia, y una vez medido esto enfocar a las áreas comerciales en estrategias de retención para los clientes, tener en claro que lograr fidelizar a un buen cliente es la mejor estrategia para un ramo como lo es automóviles.

Next Steps

El desarrollo de este proyecto es uno de los primeros pasos para la optimización dentro de la compañía, a partir de conocer cuales son esos segmentos rentables para el ramo, es decir, son apetito de riesgo para la compañía, pero que sin embargo se predicen niveles bajos de renovación se empiecen a hacer campañas de la mano con el área comercial, con el fin de fidelizar a estos clientes y poder tener carteras de negocio más solidas y rentables dentro de la compañía. Además poder iniciar con análisis de elasticidad de precios, contando con la información de la probabilidad de renovación, los análisis de benchmark y un matriz de conversión (en este caso renovación) – siniestralidad, son los primeros pasos para llegar a desarrollar análisis de optimización dentro de las carteras de la compañía, y no solamente a nivel del seguro de automóviles, sino en general a todos los seguros de la compañía especialmente del negocio de no vida, sobre el cuál el pricing, el servicio, la fidelización del cliente y la satisfacción del mismo van muy atados de la mano durante todo el desarrollo de la cobertura del seguro. Estos primeros análisis tienen mucho para aportar a las compañías en Colombia, por lo pronto como anteriormente se dijo, se iniciará el empalme con las áreas comerciales con el fin de que ellos conozcan el apetito de riesgo, y todas las oportunidades de crecimiento que tienen si se tiene un correcto análisis de los clientes. Porque como bien se ha dicho, para un agente comercial es mucho más costoso en términos de tiempos, confianza y comisiones el traer nuevos negocios a la compañía que el hacer esfuerzos de fidelizar y retener carteras conocidas y rentables, además de generar satisfacción como compañía al cliente.

7. BIBLIOGRAFÍA

Alemar Padilla, Catalina Bolance, Montserrat Guillen (2016). *Cuantificación del riesgo para la tarificación en seguros de automóvil.*

Leo Guelman, Montserrat Guillen (2013). *A causal inference approach to measure Price elasticity in automobile insurance.*

Noa Willys (2018). *Customer Satisfaction, Switching Cost and Customer Loyalty: An empirical study on the mobile telecommunication service.*

Sara Delfina Rosa Pierrend (2020) *La fidelización del cliente y retención del cliente: Tendencia que se exige hoy en día.*

Juan Felipe Leon Giraldo (2017). *Aplicación de un modelo lineal generalizado para el análisis de la retención del ramo de pesados individual en el sector asegurador.*

Measuring Performance https://topepo.github.io/caret/measuring-performance.html

Unidad 7 Modelos Lineales Generalizados
https://bookdown.org/j_morales/librostat/glm.html

Unidad 8 GLM Respuesta Binomial
https://bookdown.org/j_morales/librostat/glm.html

Epidat 4: Ayuda de Distribuciones de probabilidad. (2014) Pitat
https://www.sergas.es/Saude-publica/Documents/1899/Ayuda_Epidat_4_Distribuciones_de_probabilidad_Octubre2014.pdf

Arroyo Indira, Bravo Luis, Llinas Humberto, Muñoz Fabian (2014) *Distribuciones Poisson y Gamma: Una discreta y Continua Relación*

Salomón Micael, Carranza Juan, Piumetto Mario, Monzani Federico, Montenegro Marzo, Cordoba Augusto (2018) *Random Forest como técnica de valuación masiva del valor del suelo urbano: una aplicación para la ciudad del río cuarto, Cordoba, Argentina.*

Leo Breiman (2001) *Random Forest*

MAPFRE Colombia (2020) *Estados financieros separados al 31 diciembre de 2020 y 2019.*

La matriz de confusión y sus métricas https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Godinez Milton (2019) *Proyecto Final – Fiabilidad*

GLM – Introducción, Master en Ciencias Actuariales y Financiera. Universidad de Valencia.

Leo Breiman (1996) *Bagging Predictors*

García, J. L., Chagolla, H., & Noriega, S. (2015). *Modelos: efectos de la colinealidad en el modelado de regresión y su solución. Cultura Científica y Tecnológica*

Spiteri, M., & Azzopardi, G. (2018). Customer churn prediction for a motor insurance company.

Mark Goldburd, Anand Khare, Dan Tevet, Dmitry Guller (2019) *Generalized Linear Models for Insurance Rating.*

Francisco Folk (2022) *Modelización de la probabilidad de no renovación para una cartera de Salud.*

Circular Externa 021 de 2020.
https://www.superfinanciera.gov.co/descargas/institucional/pubFile1046025/ce021_20.docx

8. ANEXO CODIGO DE R

```
options(scipen=999)

library(tidyverse)

library(lubridate)

library(stringr)

library(writexl)

library(openxlsx)

library(readr)

library(readxl)

library(dplyr)

library(stats)

library(tidyselect)

library(DataExplorer)

library(ggplot2)

library(corrplot)

library(rcompanion)

library(RColorBrewer)

library(lattice)

library(caret)

library(randomForest)

library(tidyr)


clear_all = TRUE

if(clear_all){

  rm(list = ls())

}

gc()
```

```
# Dim_producto ------------------------------------------------------------

TAB_PROD <- read.csv("G:/Proyectos/Modelo de Retencion/DIM_PRODUCTO_MP.csv", sep = ';') %>%
  mutate(SK_PRODUCTO = as.integer(ï..SK_PRODUCTO)) %>% select(SK_PRODUCTO, COD_PRODUCTO, DESC_PRODUCTO, DESC_RAMO) %>%
  filter(str_detect(DESC_RAMO, 'AUT') | str_detect(DESC_RAMO, 'MOT')) %>%
  filter(COD_PRODUCTO != 190) %>% filter(COD_PRODUCTO != 191) %>%
  arrange(SK_PRODUCTO)



# COD_DANE ----------------------------------------------------------------
COD_DANE <- read.xlsx("G:/Proyectos/Modelo de Retencion/CODIGOS_DANE.xlsx")
# Fasecolda ---------------------------------------------------------------

GUIA_FASECOLDA <- read_excel("G:/Proyectos/Fasecolda/2023/Guía_319.xlsx", sheet = "Codigos")

GUIA_REN <- GUIA_FASECOLDA %>%
  mutate(FASECOLDA = as.character(`Codigo TXT`)) %>%
  rename(TIPO_CAJA = TipoCaja) %>%
  mutate(PESO_POTENCIA = round(as.numeric(`Peso/Potencia`),2)) %>%
  rename(RANGO_PESO_POTENCIA = `Rango_Peso/Potencia`) %>%
  select(FASECOLDA, TIPO_CAJA, PESO_POTENCIA, RANGO_PESO_POTENCIA)
# HOMOLOGACIONES ----------------------------------------------------------

RC <- read.xlsx("G:/Proyectos/Modelo de Retencion/HOMOLOGAR.xlsx", sheet = "RC")
```

```
REGIONAL <- read.xlsx("G:/Proyectos/Modelo de Retencion/HOMOLOGAR.xlsx", sheet = "Regional")

FIDELIZACION <- read.xlsx("G:/Proyectos/Modelo de Retencion/HOMOLOGAR.xlsx", sheet = "Fidelizacion")

SCORE_PERSONA <- read.xlsx("G:/Proyectos/Modelo de Retencion/Base_Score.xlsx") %>%
  mutate(ID_FEC = str_c(ID, FEC_RENOV, sep = "_")) %>%
  rename(TIPO_PERSONA = NIT) %>%
  select(ID_FEC, TIPO_PERSONA, SCORE) %>%
  unique.data.frame()



# Base 1 Prerenovaciones ----------------------------------------------------



setwd("G:/Proyectos/Modelo de Retencion/Prerenovaciones/Segunda_Version/")

file.list <- list.files(path = "G:/Proyectos/Modelo de Retencion/Prerenovaciones/Segunda_Version/", pattern = "*.csv")


Consolidado <- list()


for(i in 1:length(file.list)){
  Base <- read_delim(file.list[i], delim = ";", col_types = cols(.default = "c"))
  Consolidado[[i]] <- Base
}


Prerenovaciones <- bind_rows(Consolidado)
```

```
## LIMPIAR BASE DE PRERENOVACIONES


Renovaciones_Cleaned <- Prerenovaciones %>%
  filter(is.na(`Txt. Error`))  %>%
  filter(!is.na(`Fecha Inicio Vigencia P-R.`)) %>%
  mutate(`Poliza Individual P-R.` = str_replace(string = `Poliza Individual P-R.`,
                                                pattern = ",", replacement = ".")) %>%
  mutate(`Poliza Individual P-R.` = as.numeric(`Poliza Individual P-R.`)) %>%
  mutate(`Poliza Individual` = str_replace(string = `Poliza Individual`,
                                           pattern = ",", replacement = ".")) %>%
  mutate(NUM_POLIZA = as.numeric(`Poliza Individual`)) %>%
  mutate(CHASIS = as.character(Chasis)) %>%
  mutate(PLACA = as.character(Placa)) %>%
  select(Producto, NUM_POLIZA, `Fecha Inicio Vigencia`, `Fecha Fin Vigencia`, `Fecha
Inicio Vigencia P-R.`,
         `Fecha Fin Vigencia P-R.`, `Identificacion Asegurado`, `Nombre Asegurado`,
Nom.Regional, Nom.Sucursal, Nom.oficina, Agente,
         Fasecolda, `Grupo Tecnico`, Modelo, Marca, Linea, `Ciudad Circulacion`,
`Nombre Ciudad Circulacion`,
         Placa, CHASIS, PLACA, `Txt Rc`, `Valor Asegurado`, `Valor Asegurado P-R.`,
`Valor a Nuevo`, `Valor a Nuevo P-R.`, Siniestros,
         `Categoria Fidelizacion`, `Categoria Fidelizacion P-R.`, `Anios sin sini`, `Anios sin
sini P-R.`,
         `Dcto. Comercial Importe`, `Prima Neta Definitiva`,  `Total Comision`, `Prima
Neta Definitiva P-R.`, `Total Comision P-R.`) %>%
  mutate(`Fecha Inicio Vigencia P-R.` = dmy(`Fecha Inicio Vigencia P-R.`)) %>%
  mutate(`Fecha Inicio Vigencia` = dmy(`Fecha Inicio Vigencia`)) %>%
  mutate(`Fecha Fin Vigencia P-R.` = dmy(`Fecha Fin Vigencia P-R.`)) %>%
  mutate(`Fecha Fin Vigencia` = dmy(`Fecha Fin Vigencia`)) %>%
```

```
  mutate(`Valor Asegurado P-R.` = as.integer(`Valor Asegurado P-R.`)) %>%

  mutate(`Valor Asegurado` = as.integer(`Valor Asegurado`)) %>%

  mutate(Modelo = as.integer((Modelo))) %>%

  mutate(Siniestros = as.integer(Siniestros)) %>%

  mutate(`Valor Asegurado P-R.` = if_else(is.na(`Valor Asegurado P-R.`), `Valor Asegurado`, `Valor Asegurado P-R.`)) %>%

  mutate(`Valor a Nuevo P-R.` = if_else(is.na(`Valor a Nuevo P-R.`), `Valor a Nuevo`, `Valor a Nuevo P-R.`)) %>%

  mutate(`Categoria Fidelizacion` = if_else(is.na(`Categoria Fidelizacion`), 'Sin_Fidelizacion', `Categoria Fidelizacion`)) %>%

  mutate(`Categoria Fidelizacion P-R.` = if_else(is.na(`Categoria Fidelizacion P-R.`), 'Sin_Fidelizacion', `Categoria Fidelizacion P-R.`)) %>%

  mutate(`Anios sin sini` = as.double(`Anios sin sini`)) %>%

  mutate(`Anios sin sini` = if_else(is.na(`Anios sin sini`), 0,`Anios sin sini`)) %>%

  mutate(`Anios sin sini P-R.` = as.double(`Anios sin sini P-R.`)) %>%

  mutate(`Anios sin sini P-R.` = if_else(is.na(`Anios sin sini P-R.`), `Anios sin sini` ,`Anios sin sini P-R.`))  %>%

  mutate(across(starts_with("Prima"), ~str_replace(.x ,pattern = ",", replacement = "."))) %>%

  mutate(across(starts_with("Total Comision"), ~str_replace(.x ,pattern = ",", replacement = "."))) %>%

  mutate(`Dcto. Comercial Importe` = as.integer(`Dcto. Comercial Importe`)) %>%

  mutate(`Prima Neta Definitiva` = as.integer(`Prima Neta Definitiva`)) %>%

  mutate(`Total Comision` = as.integer(`Total Comision`)) %>%

  mutate(`Prima Neta Definitiva P-R.` = as.integer(`Prima Neta Definitiva P-R.`)) %>%

  mutate(`Total Comision P-R.` = as.integer(`Total Comision P-R.`)) %>%

  mutate(Fecha_Truncada =floor_date(`Fecha Inicio Vigencia P-R.`, unit = "month")) %>%

  mutate(Fecha_Renov = year(Fecha_Truncada)*100 + month(Fecha_Truncada)) %>%
```

```
  mutate(llave = str_c(CHASIS, PLACA, Fecha_Renov, sep = "_")) %>%

  filter(Agente != 5902) %>%

  filter(!is.na(`Prima Neta Definitiva P-R.`)) %>%

  mutate(Tip_Negocio = ifelse(Producto == '126', 'Chevy', ifelse(Producto == '131',
'Renault', ifelse(Producto == '132', 'Toyota', ifelse(Producto == '134', 'Motos',
'Individual'))))) %>%

  filter(`Grupo Tecnico` == '1' | `Grupo Tecnico` == '2' | `Grupo Tecnico` == '3') %>%

  mutate(Modelo_anios = year(Fecha_Truncada) - Modelo) %>%

  mutate(Antiguedad = if_else(Modelo_anios < 2, '-1 a 1 años',
                              if_else(Modelo_anios > 10, 'Mas de 10 años',
                                      if_else(Modelo_anios <= 5, 'Entre 1 a 5 años',
                                              'Entre 5 a 10 años')))) %>%

  mutate(Cod_Ciudad = `Ciudad Circulacion`) %>%

  select(-`Ciudad Circulacion`) %>%

  mutate(POL_FEC = str_c(NUM_POLIZA, Fecha_Renov, sep = "_")) %>%

  mutate(Fasecolda = ifelse(nchar(Fasecolda) == 8, Fasecolda,
                            paste0("0", Fasecolda)))



Renovaciones_Def <- left_join(Renovaciones_Cleaned, COD_DANE, by =
"Cod_Ciudad") %>%

  select(-COD.DEPARTAMENTO, -Código.Municipio, -CIUDAD, -COD_MUN, -
COD_DEPTO, -LLAVE.6.DIG)



colSums(is.na(Renovaciones_Cleaned))
```

```
Renovaciones_Def %>% group_by(Fecha_Renov) %>% summarise(n = n(), PRERENOV = sum(`Prima Neta Definitiva P-R.`) , ANTERIOR = sum(`Prima Neta Definitiva`),

PPROM_PRERENOV = PRERENOV/n, PPROM_ANT = ANTERIOR/n , VAR = PPROM_PRERENOV/PPROM_ANT - 1) %>% arrange(desc = n()) %>% view()


# Base 2 Seg Ventas ----------

setwd("G:/Proyectos/Modelo de Retencion/Seguimiento de Ventas/Segunda_Version/")

file.list <- list.files(path = "G:/Proyectos/Modelo de Retencion/Seguimiento de Ventas/Segunda_Version/", pattern = "*.csv")


Consolidado <- list()

for(i in 1:length(file.list)){

  Base <- read_delim(file.list[i], delim = ";", col_types = cols(.default = "c"))

  Consolidado[[i]] <- Base

}


Seg_Ventas <- bind_rows(Consolidado)


Seg_Ventas_cleaned <- Seg_Ventas %>% mutate(SK_PRODUCTO = as.integer(SK_PRODUCTO)) %>%

  mutate(NUM_POLIZA = str_replace(string = NUM_POLIZA, pattern = ",", replacement = ".")) %>%

  mutate(NUM_POLIZA = as.numeric(NUM_POLIZA))  %>%

  mutate(CLASE_SPTO = as.character(CLASE_SPTO)) %>%

  mutate(CANT_POLIZAS = as.integer(CANT_POLIZAS)) %>%

  mutate(IMP_PRIMA_ANUL = as.integer(IMP_PRIMA_NETA)) %>%
```

```
  mutate(FEC_INI_VIG_POLIZA = dmy(FEC_INI_VIG_POLIZA)) %>%
  mutate(FEC_FIN_VIG_POLIZA = dmy(FEC_FIN_VIG_POLIZA)) %>%
  mutate(FEC_ANULACION = dmy(FEC_INI_VIG_SPTO)) %>%
  mutate(NUM_SPTO = as.integer(NUM_SPTO))

Anulaciones <- left_join(Seg_Ventas_cleaned, TAB_PROD, by = "SK_PRODUCTO") %>%
  mutate(Fecha_Truncada = floor_date(FEC_INI_VIG_POLIZA, unit = "month")) %>%
  mutate(Fecha_Renov = year(Fecha_Truncada)*100 + month(Fecha_Truncada)) %>%
  mutate(POL_FEC = str_c(NUM_POLIZA, Fecha_Renov, sep = "_")) %>%
  select(COD_PRODUCTO, NUM_POLIZA, CLASE_SPTO, IMP_PRIMA_ANUL, NUM_SPTO, FEC_INI_VIG_POLIZA, FEC_FIN_VIG_POLIZA, FEC_ANULACION, Fecha_Truncada, POL_FEC) %>%
  filter(str_detect(CLASE_SPTO, 'ANUL')) %>%
  filter(!is.na(COD_PRODUCTO)) %>%
  unique.data.frame() %>%
  mutate(IMP_PRIMA_ANUL = ifelse(is.na(IMP_PRIMA_ANUL), 0, IMP_PRIMA_ANUL))


colSums(is.na(Anulaciones))


# Data Emision ------------------------------------------------------------
  FEC_EMISION <- read.csv("G:/Proyectos/Modelo de Retencion/Fecha_emision.csv", sep = ";") %>%
    select(NUM_POLIZA.x, FECHA_EMISION) %>%
    mutate(FECHA_EMISION = dmy(FECHA_EMISION)) %>%
```

```
  unique.data.frame()



# Ajuste Suavización ----------------------------------------------------

setwd("G:/Proyectos/Modelo de Retencion/Ajuste_Suavizacion/Suavizacion/")

file.list <- list.files(path = "G:/Proyectos/Modelo de Retencion/Ajuste_Suavizacion/Suavizacion/", pattern = "*.csv")


Consolidado <- list()


for(i in 1:length(file.list)){
 Base <- read_delim(file.list[i], delim = ";", col_types = cols(.default = "c"))
 Consolidado[[i]] <- Base
}


Suavizacion <- bind_rows(Consolidado) %>%
 filter(!is.na(NOM_FEC)) %>%
 unique.data.frame()


Suav <- Suavizacion %>%
 mutate(AJUSTE_SUAV = as.integer(IMPORTE)) %>%
 select(NOM_FEC, AJUSTE_SUAV)


Suavizacion %>%  group_by(FEC_RENOV) %>%  summarise(n = n()) %>%  view()

# Vigentes ---------------------------------------------------------------
```

```
Vigentes <- read.csv("G:/Proyectos/Modelo de Retencion/Vigentes/Resumen/Vigentes_resumen_202101_202302.csv", sep = ";")

Vigentes <- Vigentes %>%

 rename(ind = X) %>%

 mutate(PLACA = C__en__COD_PLACA) %>%

 mutate(CHASIS = COD_CHASIS_DV) %>%

 mutate(RCE = C__en__NUM_OPCION_RC_DV) %>%

 mutate(ANIOS_SIN_SINI = C__en__NUM_ANIO_NOSINI) %>%

 mutate(PRIMA_NETA = as.integer(C__en__IMP_PRIMA_ANUAL)) %>%

 mutate(Fecha_Inicio = Fecha_lla) %>%

 mutate(FEC_INI_VIG_RIESGO = dmy(FEC_INI_VIG_RIESGO)) %>%

 mutate(FEC_FIN_VIG_RIESGO = dmy(FEC_FIN_VIG_RIESGO)) %>%

 mutate(FEC_NACIMIENTO_ASEG = dmy(FEC_NACIMIENTO_ASEG)) %>%

 mutate(EDAD = as.numeric(round((FEC_INI_VIG_RIESGO - FEC_NACIMIENTO_ASEG)/365,0))) %>%

 select(ind, COD_PRODUCTO, NUM_POLIZA, FEC_INI_VIG_RIESGO, FEC_FIN_VIG_RIESGO,Fecha_Inicio,

     NOM_DEPARTAMENTO, PLACA, CHASIS, FEC_NACIMIENTO_ASEG, EDAD, GENERO, PRIMA_NETA, llave)



colSums(is.na(Vigentes))



# Left Join entre Prerenovaciones  y Vigentes ----------------------------

Pre_Vig <- left_join(Renovaciones_Def, Vigentes, by = "llave")

Pre_Vig <- Pre_Vig %>%  mutate(NUM_POLIZA = NUM_POLIZA.x) %>%
```

```
  mutate(Mca_vig = if_else(is.na(ind), 0, 1)) %>%

  unique.data.frame()



Resumen <- Pre_Vig %>%  group_by(llave) %>%  summarise(n = n()) %>% arrange(desc(n)) %>%  view()


  #TIPO NEGOCIO

  TAB1 <-  table(TIP_NEGOCIO = Pre_Vig$Tip_Negocio) %>%  view()

  TAB2 <-  table(TIP_NEGOCIO = Pre_Vig$Tip_Negocio, Pre_Vig$Mca_vig) %>% view()

  TAB_TIP_NEG <- left_join(TAB2, TAB1, by = 'TIP_NEGOCIO') %>% mutate(INDICE = Freq.x/Freq.y) %>%  arrange(desc(Freq.y)) %>%  filter(Var2 == 1) %>% view()


  #FECHA RENOVACIÓN

  TAB1.1 <-  table(FEC_RENOV = Pre_Vig$Fecha_Renov) %>%  view()

  TAB2.1 <-  table(FEC_RENOV = Pre_Vig$Fecha_Renov, Pre_Vig$Mca_vig) %>% view()

  TAB_FEC_RENOV <- left_join(TAB2.1, TAB1.1, by = 'FEC_RENOV') %>% mutate(INDICE = Freq.x/Freq.y) %>%  arrange(FEC_RENOV) %>%  filter(Var2 == 1) %>% view()



Data0 <- left_join(Pre_Vig, Anulaciones, by = "POL_FEC" )

Data0 <- Data0 %>%  mutate(Altura_Anul = as.numeric(round((FEC_ANULACION - FEC_INI_VIG_RIESGO)/365, 2))) %>%

  mutate(Mca_Anul = ifelse(Altura_Anul > 0, 1 ,

                           ifelse(Altura_Anul < 0 | CLASE_SPTO == "ANULACIONP", 1, 0)))
```

```
Data0$Mca_Anul[is.na(Data0$Mca_Anul)] = 1

Data0 <- Data0 %>%  mutate(Mca_renov = Mca_vig * Mca_Anul) %>%

  mutate(RC_LIM0 =`Txt Rc`) %>%

  mutate(FIDELIZACION0 = `Categoria Fidelizacion P-R.`) %>%

  mutate(REGIONAL0 = Nom.Regional) %>%

  select(-`Fecha Inicio Vigencia`, - `Fecha Fin Vigencia`, -`Nombre Ciudad Circulacion`, -Placa, -`Valor Asegurado`,

       -`Valor a Nuevo`, -`Categoria Fidelizacion`, -Fecha_Truncada.x, -ind, -NOM_DEPARTAMENTO, -PLACA.y, -CHASIS.y,

       -COD_PRODUCTO.y, -COD_PRODUCTO.x, -FEC_INI_VIG_POLIZA, -FEC_FIN_VIG_POLIZA, -Fecha_Truncada.y,

        -NUM_POLIZA.y, -NUM_POLIZA.x.x, -NUM_POLIZA.y.y, -POL_FEC, -`Txt Rc`, -`Categoria Fidelizacion P-R.`, -Nom.Regional)


Data1 <- left_join(Data0, RC, by = "RC_LIM0")

Data2 <- left_join(Data1, REGIONAL, by = "REGIONAL0")

Data3 <- left_join(Data2, FIDELIZACION, by = "FIDELIZACION0")



Data4 <- left_join(Data3, FEC_EMISION, by = "NUM_POLIZA.x") %>%

  mutate(ALTURA_RENOV = as.numeric(round((`Fecha Inicio Vigencia P-R.` - FECHA_EMISION)/365, 0))) %>%

  select(-RC_LIM0, -FIDELIZACION0, -REGIONAL0, - NUM_SPTO) %>%

  rename(PRODUCTO = Producto, NUM_POLIZA = NUM_POLIZA.x, FECHA_INICIO = `Fecha Inicio Vigencia P-R.`, FECHA_FIN = `Fecha Fin Vigencia P-R.`,

       ID = `Identificacion Asegurado`,SUCURSAL = Nom.Sucursal, OFICINA = Nom.oficina, AGENTE = Agente, FASECOLDA = Fasecolda, TIPO_VEHICULO = `Grupo Tecnico`,
```

```
    MODELO = Modelo, MARCA = Marca, LINEA = Linea, CHASIS = CHASIS.x, PLACA
= PLACA.x, VA_COMERCIAL = `Valor Asegurado P-R.`, VA_NUEVO = `Valor a Nuevo
P-R.`,

    SINIESTROS = Siniestros, BONUS_MALUS = `Anios sin sini P-R.`,
IMP_DESC_COMERCIAL = `Dcto. Comercial Importe`, PRIMA_ANTERIOR = `Prima
Neta Definitiva`,

    COMISION_ANTERIOR = `Total Comision`, PRIMA_PRERENOV = `Prima Neta
Definitiva P-R.`, COMISION_PRERENOV = `Total Comision P-R.`, FEC_RENOV =
Fecha_Renov,

    NEGOCIO = Tip_Negocio, ANTIGUEDAD = Modelo_anios,
RANGO_ANTIGUEDAD = Antiguedad, PRIMA_RENOV = PRIMA_NETA, MCA_VIG =
Mca_vig, ALTURA_ANUL = Altura_Anul,

    MCA_ANUL = Mca_Anul, MCA_RENOV = Mca_renov, ASEGURADO = `Nombre
Asegurado`) %>%

  unique.data.frame() %>%

  filter(ALTURA_RENOV != 0) %>%

  filter(EDAD > 0 | is.na(EDAD)) %>%

  mutate(RANGO_PRIMA_ANT = ifelse(PRIMA_ANTERIOR <= 1000000, "1. Inf a 1
millon",
                                  ifelse(PRIMA_ANTERIOR >= 1500000, "3. Mayor a 1.5 millones",

                                         "2. Entre 1 y 1.5 millones"))) %>%

  mutate(ID_FEC = str_c(ID, FEC_RENOV, sep = "_")) %>%

  mutate(NOM_FEC = str_c(NUM_POLIZA, FEC_RENOV, sep = "_"))



Data5 <- left_join(Data4, GUIA_REN, by = "FASECOLDA")

Data6 <- left_join(Data5, SCORE_PERSONA, by = "ID_FEC")

Data7 <- left_join(Data6, Suav, by = "NOM_FEC") %>%

  mutate(SUAVIZACION = ifelse(is.na(AJUSTE_SUAV), 0, AJUSTE_SUAV)) %>%

  mutate(PRIMA_EMBLEM = PRIMA_PRERENOV - SUAVIZACION) %>%
```

```
  mutate(SCORE = as.integer(SCORE))


Vector <- Data4 %>%  group_by(llave) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%  filter(n >= 2) %>%
  unique.data.frame()


Data <- left_join(Data7, Vector, by = "llave")  %>%
  filter(is.na(n)) %>%
  select(-n, -FEC_INI_VIG_RIESGO, -FEC_FIN_VIG_RIESGO, - Fecha_Inicio, -`Anios sin
sini`, -PRIMA_RENOV,
         -ID_FEC, -NOM_FEC, -AJUSTE_SUAV) %>%
  mutate(VAR_PRIMA = round((PRIMA_PRERENOV/PRIMA_ANTERIOR)-1,2)) %>%
  unique.data.frame() %>%
  mutate(RANGO_SCORE = if_else(TIPO_PERSONA == 'NIT', "NIT",
                               if_else(SCORE < 150, as.character(SCORE),
                               if_else(between(SCORE, 150, 250), "Mayor a 150 y menor o igual a
250",
                               if_else(between(SCORE, 251, 350), "Mayor a 250 y menor o igual a
350",
                               if_else(between(SCORE, 351, 450), "Mayor a 350 y menor o igual a
450",
                               if_else(between(SCORE, 451, 600), "Mayor a 450 y menor o igual a
600",
                               if_else(between(SCORE, 601, 700), "Mayor a 600 y menor o igual a
700",
                               if_else(between(SCORE, 701, 750), "Mayor a 700 y menor o igual a
750",
                               if_else(between(SCORE, 751, 800), "Mayor a 751 y menor o igual a
800",
```

```
                    if_else(between(SCORE, 801, 850), "Mayor a 800 y menor o igual a
850",
                    if_else(between(SCORE, 851, 900), "Mayor a 850 y menor o igual a
900",
                        "Mayor a 900")))))))))))) %>%
  mutate(RANGO_EDAD = if_else(TIPO_PERSONA == 'NIT', "NIT",
                    if_else(is.na(EDAD), "1. Sin informacion",
                    if_else(EDAD <= 18, "0. Menor o igual a 18 años",
                    if_else(between(EDAD, 19, 25), "2. Mayor a 19 y menor o igual a
25",
                    if_else(between(EDAD, 26, 35), "3. Mayor a 25 y menor o igual a
35",
                    if_else(between(EDAD, 36, 45), "4. Mayor a 35 y menor o igual a
45",
                    if_else(between(EDAD, 46, 55), "5. Mayor a 45 y menor o igual a
55",
                    if_else(between(EDAD, 56, 65), "6. Mayor a 55 y menor o igual a
65",
                    if_else(between(EDAD, 66, 70), "7. Mayor a 65 y menor o igual a
70",
                    "8. Mayor a 70")))))))))) %>%
  mutate(GENERO2 = if_else(is.na(GENERO), "Indefinido",
                   if_else(TIPO_PERSONA == 'NIT', "NIT",
                   if_else(GENERO == '-', "Indefinido",
                   if_else(is.na(GENERO), "Indefinido", GENERO))))) %>%
  select(-GENERO) %>%
  rename(GENERO = GENERO2) %>%
  mutate(RANGO_VAR_PRIMA = if_else(VAR_PRIMA <= -0.05, "0. Menor o igual a -
5%",
                   if_else(between(VAR_PRIMA, -0.04, 0.00), "2. Mayor a -5% y menor
o igual a 0%",
```

```
                    if_else(between(VAR_PRIMA, 0.01, 0.05), "3. Mayor a 0% y menor o
igual a 5%",

                    if_else(between(VAR_PRIMA, 0.06, 0.12), "4. Mayor a 5% y menor o
igual a 12%",

                    if_else(between(VAR_PRIMA, 0.12, 0.18), "5. Mayor a 12% y menor
o igual a 18%",

                    if_else(between(VAR_PRIMA, 0.18, 0.24), "6. Mayor a 18% y menor
o igual a 24%",

                    if_else(between(VAR_PRIMA, 0.24, 0.30), "7. Mayor a 24% y menor
o igual a 30%",

                    "8. Mayor a 30%")))))))) %>%

 mutate(RANGO_PESO_POTENCIA = ifelse(is.na(RANGO_PESO_POTENCIA), "Sin-
Info", RANGO_PESO_POTENCIA)) %>%

 mutate(TIPO_CAJA = ifelse(is.na(TIPO_CAJA), "Sin_Info", TIPO_CAJA)) %>%

 select(-ASEGURADO) %>%

 mutate(SCORE = as.character(ifelse(is.na(SCORE), -1, SCORE)))


colSums(is.na(Data))



# Distribución de Variables ----------------------------------------------


# TIPO NEGOCIO

TAB1.2 <-  table(NEGOCIO = Data$NEGOCIO) %>%  view()

TAB2.2 <-  table(NEGOCIO = Data$NEGOCIO, Data$MCA_RENOV) %>%  view()

TAB_TIP_NEG1 <- left_join(TAB2.2, TAB1.2, by = 'NEGOCIO') %>% mutate(INDICE =
Freq.x/Freq.y) %>%

 arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_TIP_NEG1$INDICE,  names  =  TAB_TIP_NEG1$NEGOCIO,     ylab  =  '%
Renovación', las =  2, col = brewer.pal(n = 11, name = "RdBu"),
```

```
      main ="Indice de Renovación - Tipo Negocio")

barplot(prop.table(table(Data$NEGOCIO)),  ylab = "Proportion",las =  2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia de Variable - Tipo Negocio")


## RENOVACION

TAB1.3 <-  table(FEC_RENOVACION = Data$FEC_RENOV) %>%  view()

TAB2.3 <-  table(FEC_RENOVACION = Data$FEC_RENOV, Data$MCA_RENOV) %>% view()

TAB_FEC_RENOV3  <-  left_join(TAB2.3,  TAB1.3,  by  =  'FEC_RENOVACION')  %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(FEC_RENOVACION)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_FEC_RENOV3$INDICE, names = TAB_FEC_RENOV3$FEC_RENOVACION, ylab = '% Renovación', las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Fecha de Renovación")

barplot(prop.table(table(Data$FEC_RENOV)),  ylab = "Proportion", las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia de Variable - Fecha de Renovación")


## MARCA

TAB1.4 <-  table(MARCA = Data$MARCA) %>%  view()

TAB2.4 <-  table(MARCA = Data$MARCA, Data$MCA_RENOV) %>%  view()

TAB_MARCA  <-  left_join(TAB2.4,  TAB1.4,  by  =  'MARCA')  %>%  mutate(INDICE  = Freq.x/Freq.y) %>%

  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>% head(20) %>% view()

barplot(TAB_MARCA$INDICE, names = TAB_MARCA$MARCA, ylab = '% Renovación', las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Marca")

barplot(prop.table(table(Data$MARCA)),  ylab  =  "Proportion",las  =  2,  col  = brewer.pal(n = 11, name = "RdBu"),
```

```
      main ="Frecuencia - Marca")


## GRUPO TECNICO

TAB1.5 <-  table(TIPO_VEHICULO = Data$TIPO_VEHICULO) %>%  view()

TAB2.5 <-  table(TIPO_VEHICULO = Data$TIPO_VEHICULO, Data$MCA_RENOV) %>%
view()

TAB_GTECNICO  <-  left_join(TAB2.5,  TAB1.5,  by  =  'TIPO_VEHICULO')  %>%
mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_GTECNICO$INDICE,  names  =  TAB_GTECNICO$TIPO_VEHICULO,  xlab  =
'TIPO VEHICULO', ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Grupo Técnico")

barplot(prop.table(table(Data$TIPO_VEHICULO)),  xlab  =  "TIPO_VEHICULO",  ylab  =
"Proportion", las = 1,

      col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Grupo Técnico")


## REGIONAL

TAB1.6 <-  table(REGIONAL = Data$REGIONAL) %>%  view()

TAB2.6 <-  table(REGIONAL = Data$REGIONAL, Data$MCA_RENOV) %>%  view()

TAB_REGIONAL <- left_join(TAB2.6, TAB1.6, by = 'REGIONAL') %>% mutate(INDICE =
Freq.x/Freq.y) %>%

  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_REGIONAL$INDICE,  names  =  TAB_REGIONAL$REGIONAL,  ylab  =  '%
Renovación',

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Regional",

      cex.names = 0.4)
```

```
barplot(prop.table(table(Data$REGIONAL)),  ylab = "Proportion",
        las = 2, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Frecuencia - Regional",
        cex.names = 0.4)


## PRODUCTO
TAB1.7 <-  table(PRODUCTO = Data$PRODUCTO) %>%  view()
TAB2.7 <-  table(PRODUCTO = Data$PRODUCTO, Data$MCA_RENOV) %>%  view()
TAB_PROD <- left_join(TAB2.7, TAB1.7, by = 'PRODUCTO') %>% mutate(INDICE = Freq.x/Freq.y) %>%
  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()
barplot(TAB_PROD$INDICE, names = TAB_PROD$PRODUCTO, xlab = 'Producto', ylab = '% Renovación',
        las = 2, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Indice de Renovación - Producto")
barplot(prop.table(table(Data$PRODUCTO)), xlab = "PRODUCTO", ylab = "Proportion",
        las = 2, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Frecuencia - Producto")


## SINIESTROS
TAB1.8 <-  table(N_SINI = Data$SINIESTROS) %>%  view()
TAB2.8 <-  table(N_SINI = Data$SINIESTROS, Data$MCA_RENOV) %>%  view()
TAB_NSINI <- left_join(TAB2.8, TAB1.8, by = 'N_SINI') %>% mutate(INDICE = Freq.x/Freq.y) %>%
  arrange(N_SINI)  %>%  filter(Var2 == 1) %>%  view()
barplot(TAB_NSINI$INDICE, names = TAB_NSINI$N_SINI, xlab = 'Numéro de Siniestros', ylab = '% Renovación',
        las = 1, col = brewer.pal(n = 11, name = "RdBu"),
```

```
      main ="Indice de Renovación -  Cantidad de Siniestros")

barplot(prop.table(table(Data$SINIESTROS)), xlab = "SINIESTROS", ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Cantidad de Siniestros")


## ANIOS SIN SINIESTROS

TAB1.9 <-  table(BONUS_MALUS = Data$BONUS_MALUS) %>%  view()

TAB2.9 <-  table(BONUS_MALUS = Data$BONUS_MALUS, Data$MCA_RENOV) %>% view()

TAB_BONUS <- left_join(TAB2.9, TAB1.9, by = 'BONUS_MALUS') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(BONUS_MALUS)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_BONUS$INDICE, names = TAB_BONUS$BONUS_MALUS, xlab = 'Años sin siniestro', ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Años sin Siniestro")

barplot(prop.table(table(Data$BONUS_MALUS)), xlab = "BONUS MALUS", ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Años sin Siniestro")


## ALTURA RENOV

TAB1.10 <-  table(ALTURA_RENOV = Data$ALTURA_RENOV) %>%  view()

TAB2.10 <-  table(ALTURA_RENOV = Data$ALTURA_RENOV, Data$MCA_RENOV) %>%  view()

TAB_ALTURA_RENOV <- left_join(TAB2.10, TAB1.10, by = 'ALTURA_RENOV') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(ALTURA_RENOV)  %>%  filter(Var2 == 1) %>%  view()
```

```
barplot(TAB_ALTURA_RENOV$INDICE, names = TAB_ALTURA_RENOV$ALTURA_RENOV, xlab = 'Altura Renovación',
        ylab = '% Renovación', las = 1, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Indice de Renovación - Altura de Renovación",
        cex.names = 0.75)
barplot(prop.table(table(Data$ALTURA_RENOV)), xlab = "ALTURA RENOVACION", ylab = "Proportion",
        las = 1, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Frecuencia - Altura de Renovación",
        cex.names = 0.75)


# FIDELIZACION
TAB1.11 <-  table(FIDELIZACION = Data$FIDELIZACION) %>%  view()
TAB2.11 <-   table(FIDELIZACION = Data$FIDELIZACION, Data$MCA_RENOV) %>% view()
TAB_FIDELIZACION <- left_join(TAB2.11, TAB1.11, by = 'FIDELIZACION') %>% mutate(INDICE = Freq.x/Freq.y) %>%
  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()
barplot(TAB_FIDELIZACION$INDICE, names = TAB_FIDELIZACION$FIDELIZACION, ylab = '% Renovación',
        las = 1, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Indice de Renovación - Categoría de Fidelización",
        cex.names = 0.75)
barplot(prop.table(table(Data$FIDELIZACION)),  ylab = "Proportion",
        las = 1, col = brewer.pal(n = 11, name = "RdBu"),
        main ="Frecuencia - Categoría de Fidelización",
        cex.names = 0.75)


## LIMITE DE RC
```

```
TAB1.12 <-  table(RC_LIM = Data$RC) %>%  view()

TAB2.12 <-  table(RC_LIM = Data$RC, Data$MCA_RENOV) %>%  view()

TAB_RC_LIM <- left_join(TAB2.12, TAB1.12, by = 'RC_LIM') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_RC_LIM$INDICE, names = TAB_RC_LIM$RC_LIM, xlab = 'RC Limite', ylab = '% Renovación',

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Límite de RC")

barplot(prop.table(table(Data$RC)), xlab = "LIMITE DE RC", ylab = "Proportion",

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Límite de RC")


##RANGO PRIMA ANTERIOR

TAB1.13 <-  table(RANG_PRIM_ANT = Data$RANGO_PRIMA_ANT) %>%  view()

TAB2.13 <-  table(RANG_PRIM_ANT = Data$RANGO_PRIMA_ANT, Data$MCA_RENOV) %>%  view()

TAB_RANG_PANT <- left_join(TAB2.13, TAB1.13, by = 'RANG_PRIM_ANT') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(RANG_PRIM_ANT)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_RANG_PANT$INDICE, names = TAB_RANG_PANT$RANG_PRIM_ANT, ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Rango de Prima Anterior",

      cex.names = 0.75)

barplot(prop.table(table(Data$RANGO_PRIMA_ANT)),  ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Rango de Prima Anterior",

      cex.names = 0.75)
```

```
##RANGO ANTIGUEDAD VEHICULO

TAB1.14 <-  table(RANG_ANTIGUEDAD = Data$RANGO_ANTIGUEDAD) %>%  view()

TAB2.14 <-  table(RANG_ANTIGUEDAD = Data$RANGO_ANTIGUEDAD, Data$MCA_RENOV) %>%  view()

TAB_RANG_ANTIGUEDAD <-  left_join(TAB2.14, TAB1.14, by = 'RANG_ANTIGUEDAD') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(RANG_ANTIGUEDAD)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_RANG_ANTIGUEDAD$INDICE, names = TAB_RANG_ANTIGUEDAD$RANG_ANTIGUEDAD,

      ylab = '% Renovación',las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Rango de Antiguedad Vehiculo")

barplot(prop.table(table(Data$RANGO_ANTIGUEDAD)),  ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Rango de Antiguedad Vehiculo")


##EDAD

TAB1.15 <-  table(EDAD = Data$EDAD) %>%  view()

TAB2.15 <-  table(EDAD = Data$EDAD, Data$MCA_RENOV) %>%  view()

TAB_EDAD <-  left_join(TAB2.15, TAB1.15, by = 'EDAD') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(EDAD)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_EDAD$INDICE, names = TAB_EDAD$EDAD, xlab = 'Edad', ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Edad")

barplot(prop.table(table(Data$EDAD)), xlab = "EDAD", ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),
```

```
      main ="Frecuencia - Edad")


##RANGO EDAD

TAB1.16 <-  table(RANGO_EDAD = Data$RANGO_EDAD) %>%  view()

TAB2.16 <-   table(RANGO_EDAD = Data$RANGO_EDAD, Data$MCA_RENOV) %>%
view()

TAB_RANG_EDAD  <-  left_join(TAB2.16,  TAB1.16,  by  =  'RANGO_EDAD')  %>%
mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(RANGO_EDAD)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_RANG_EDAD$INDICE, names = TAB_RANG_EDAD$RANGO_EDAD, ylab
= '% Renovación',

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Rango de Edad",

      cex.names = 0.5)

barplot(prop.table(table(Data$RANGO_EDAD)),  ylab = "Proportion",

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Rango de  Edad",

      cex.names = 0.5)


##GENERO

TAB1.17 <-  table(GENERO = Data$GENERO) %>%  view()

TAB2.17 <-  table(GENERO = Data$GENERO, Data$MCA_RENOV) %>%  view()

TAB_GENERO <- left_join(TAB2.17, TAB1.17, by = 'GENERO') %>% mutate(INDICE =
Freq.x/Freq.y) %>%

  arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_GENERO$INDICE,  names  =  TAB_GENERO$GENERO,  ylab  =  '%
Renovación',

      las = 2, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Género")
```

```
barplot(prop.table(table(Data$GENERO)),  ylab = "Proportion",
      las = 2, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Frecuencia - Género")


##VARIACION PRIMA
TAB1.18 <-  table(VAR_PRIMA = Data$RANGO_VAR_PRIMA) %>%  view()
TAB2.18 <-  table(VAR_PRIMA = Data$RANGO_VAR_PRIMA, Data$MCA_RENOV) %>%  view()
TAB_VAR_PRIMA <- left_join(TAB2.18, TAB1.18, by = 'VAR_PRIMA') %>% mutate(INDICE = Freq.x/Freq.y) %>%
 arrange(VAR_PRIMA)  %>%  filter(Var2 == 1) %>%  view()
barplot(TAB_VAR_PRIMA$INDICE, names = TAB_VAR_PRIMA$VAR_PRIMA, ylab = '% Renovación',
      las = 2, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Indice de Renovación - Variación de Prima",
      cex.names = 0.5)
barplot(prop.table(table(Data$RANGO_VAR_PRIMA)),  ylab = "Proportion",
      las = 2, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Frecuencia - Variación de Prima",
      cex.names = 0.5)


##TIPO PERSONA
TAB1.19 <-  table(TIPO_PERSONA = Data$TIPO_PERSONA) %>%  view()
TAB2.19 <-  table(TIPO_PERSONA = Data$TIPO_PERSONA, Data$MCA_RENOV) %>% view()
TAB_PERSONA <- left_join(TAB2.19, TAB1.19, by = 'TIPO_PERSONA') %>% mutate(INDICE = Freq.x/Freq.y) %>%
 arrange(desc(Freq.y))  %>%  filter(Var2 == 1) %>%  view()
```

```
barplot(TAB_PERSONA$INDICE, names = TAB_PERSONA$TIPO_PERSONA, xlab = 'Tipo de persona', ylab = '% Renovación',
      las = 1, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Indice de Renovación - Tipo de Persona")
barplot(prop.table(table(Data$TIPO_PERSONA)), xlab = "TIPO DE PERSONA", ylab = "Proportion",
      las = 1, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Frecuencia - Tipo de Persona")

##SCORE
TAB1.21 <-  table(SCORE = Data$RANGO_SCORE) %>%  view()
TAB2.21 <-  table(SCORE = Data$RANGO_SCORE, Data$MCA_RENOV) %>%  view()
TAB_SCORE <- left_join(TAB2.21, TAB1.21, by = 'SCORE') %>% mutate(INDICE = Freq.x/Freq.y) %>%
  arrange(SCORE)  %>%  filter(Var2 == 1) %>% head(200) %>%  view()
barplot(TAB_SCORE$INDICE, names = TAB_SCORE$SCORE,  ylab = '% Renovación',
      las = 2, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Indice de Renovación - Rango Score Financiero",
      cex.names = 0.55)
barplot(prop.table(table(Data$RANGO_SCORE)),  ylab = "Proportion",
      las = 2, col = brewer.pal(n = 11, name = "RdBu"),
      main ="Frecuencia - Rango Score Financiero",
      cex.names = 0.5)

##RANGO PESO POTENCIA
TAB1.22 <-  table(PESO_POTENCIA = Data$RANGO_PESO_POTENCIA) %>%  view()
TAB2.22 <-  table(PESO_POTENCIA = Data$RANGO_PESO_POTENCIA, Data$MCA_RENOV) %>%  view()
```

```
TAB_PESOPOT <- left_join(TAB2.22, TAB1.22, by = 'PESO_POTENCIA') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(PESO_POTENCIA)  %>%  filter(Var2 == 1) %>%  view()

barplot(TAB_PESOPOT$INDICE, names = TAB_PESOPOT$PESO_POTENCIA, ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Rango Peso Potencia",

      cex.names = 0.5)

barplot(prop.table(table(Data$RANGO_PESO_POTENCIA)),  ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Rango Peso Potencia",

      cex.names = 0.5)


##TIPO CAJA

TAB1.23 <-  table(TIPO_CAJA = Data$TIPO_CAJA) %>%  view()

TAB2.23 <-  table(TIPO_CAJA = Data$TIPO_CAJA, Data$MCA_RENOV) %>%  view()

TAB_CAJA <- left_join(TAB2.23, TAB1.23, by = 'TIPO_CAJA') %>% mutate(INDICE = Freq.x/Freq.y) %>%

  arrange(Freq.y)  %>%  filter(Var2 == 1) %>% head(200) %>%  view()

barplot(TAB_CAJA$INDICE, names = TAB_CAJA$TIPO_CAJA, xlab = 'Tipo Caja', ylab = '% Renovación',

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Indice de Renovación - Tipo de Caja")

barplot(prop.table(table(Data$TIPO_CAJA)), xlab = "TIPO DE CAJA", ylab = "Proportion",

      las = 1, col = brewer.pal(n = 11, name = "RdBu"),

      main ="Frecuencia - Tipo de Caja")
```

```
# Variables categoricas/numericas ------------------------------------------------------------

Categoricas <- Data %>%
  select(PRODUCTO, SUCURSAL, OFICINA, AGENTE, TIPO_VEHICULO, MARCA, LINEA,
         NEGOCIO, RANGO_ANTIGUEDAD, Cod_Ciudad, DEPARTAMENTO, GENERO,REGIONAL, FIDELIZACION,
         RANGO_PRIMA_ANT, TIPO_CAJA, RANGO_PESO_POTENCIA, TIPO_PERSONA, RANGO_SCORE,
         RANGO_EDAD,RANGO_VAR_PRIMA, CLASE_SPTO, SCORE)

itera_grid = expand.grid(names(Categoricas), names(Categoricas)) %>%
 tibble

V_Cramer_tbl = pmap(
 .l = itera_grid,
 .f = ~ {
  chiq_test = chisq.test(Categoricas[[..1]], Categoricas[[..2]])
  tibble(
   var1 = ..1,
   var2 = ..2,
   coef_chi = chiq_test[["statistic"]],
   n = length(Categoricas[[..1]]),
   q = table(Categoricas[[..1]], Categoricas[[..2]]) %>%
    dim %>%
    min
    # min(table(Categoricas[[..1]], Categoricas[[..2]]))
  ) %>%
```

```
    mutate(
      V_Cramer = sqrt(coef_chi/((q-1)*n))
    )
  }
)  %>%
  bind_rows %>%
  select(var1, var2, V_Cramer) #%>%
  # pivot_wider(id_cols = var2)

V_Cramer_mtx = V_Cramer_tbl %>%
  mutate(V_Cramer = round(V_Cramer, 1)) %>%
  pivot_wider(names_from = var2, values_from = V_Cramer)  %>%
  select(-var1) %>%
  as.matrix

row.names(x = V_Cramer_mtx) = pivot_wider(data = V_Cramer_tbl, names_from = var2, values_from = V_Cramer)[["var1"]]
ggcorrplot::ggcorrplot(
  corr = V_Cramer_mtx, hc.order = TRUE, type = "lower" , lab = TRUE, title = "Correlación (V-Cramer)"
)
```

```
Numericas <- Data %>%
```

```
 select(MODELO, VA_COMERCIAL, SINIESTROS, BONUS_MALUS, IMP_DESC_COMERCIAL, PRIMA_ANTERIOR, COMISION_ANTERIOR,

      PRIMA_PRERENOV, COMISION_PRERENOV, FEC_RENOV, ALTURA_ANUL, ALTURA_RENOV,ANTIGUEDAD, RC,  PESO_POTENCIA , EDAD,

      SUAVIZACION, PRIMA_EMBLEM, VAR_PRIMA)

Num.cor <- cor(Numericas, method = "pearson", use = "complete.obs")

ggcorrplot::ggcorrplot(

 corr = Num.cor, hc.order = TRUE, type = "lower" , lab = TRUE, title = "Correlación (Pearson)"

)


ggcorrplot::corrplot(Num.cor, method = "number", type = "lower")



Data_SinCorr <- Data %>%

 select(-Cod_Ciudad, -SUCURSAL, - OFICINA,-AGENTE, -PRODUCTO, -SCORE, -TIPO_PERSONA, -EDAD, -LINEA, -PESO_POTENCIA, -MODELO,

      -COMISION_ANTERIOR, - COMISION_PRERENOV, - PRIMA_EMBLEM, -PRIMA_ANTERIOR, - RANGO_PESO_POTENCIA, - RANGO_EDAD,

      -TIPO_CAJA, -VA_COMERCIAL)


 ## write.xlsx(Data_SinCorr, "G://Proyectos/Modelo de Retencion/SALIDA/Data_SinCrr.xlsx", overwrite = TRUE)


# Categoricas Sin Correlación ------------------------------------------


Categoricas_2 <- Data_SinCorr %>%

 select(TIPO_VEHICULO, MARCA, NEGOCIO, RANGO_ANTIGUEDAD, DEPARTAMENTO, GENERO,REGIONAL, FIDELIZACION,

      RANGO_PRIMA_ANT,  RANGO_SCORE,RANGO_VAR_PRIMA, CLASE_SPTO)
```

```
itera_grid = expand.grid(names(Categoricas_2), names(Categoricas_2)) %>%
  tibble


V_Cramer_tbl = pmap(
  .l = itera_grid,
  .f = ~ {
    chiq_test = chisq.test(Categoricas_2[[..1]], Categoricas_2[[..2]])
    tibble(
      var1 = ..1,
      var2 = ..2,
      coef_chi = chiq_test[["statistic"]],
      n = length(Categoricas_2[[..1]]),
      q = table(Categoricas_2[[..1]], Categoricas_2[[..2]]) %>%
        dim %>%
        min
      # min(table(Categoricas[[..1]], Categoricas[[..2]]))
    ) %>%
      mutate(
        V_Cramer = sqrt(coef_chi/((q-1)*n))
      )
  }
)  %>%
  bind_rows %>%
  select(var1, var2, V_Cramer) #%>%
# pivot_wider(id_cols = var2)
```

```
V_Cramer_mtx = V_Cramer_tbl %>%
  mutate(V_Cramer = round(V_Cramer, 1)) %>%
  pivot_wider(names_from = var2, values_from = V_Cramer)  %>%
  select(-var1) %>%
  as.matrix

row.names(x = V_Cramer_mtx) = pivot_wider(data = V_Cramer_tbl, names_from = var2, values_from = V_Cramer)[["var1"]]
ggcorrplot::ggcorrplot(
  corr = V_Cramer_mtx, hc.order = TRUE, type = "lower" , lab = TRUE, title = "Correlación (V-Cramer)"
)
```

```
# Númericas Sin Correlación ----------------------------------------------
Numericas_2 <- Data_SinCorr %>%
  select(SINIESTROS, BONUS_MALUS, IMP_DESC_COMERCIAL,
       PRIMA_PRERENOV, FEC_RENOV, ALTURA_ANUL, ALTURA_RENOV,ANTIGUEDAD, RC,
       SUAVIZACION, VAR_PRIMA)
Num.cor_2 <- cor(Numericas_2, method = "pearson", use = "complete.obs")
ggcorrplot::ggcorrplot(
  corr = Num.cor_2, hc.order = TRUE, type = "lower" , lab = TRUE, title = "Correlación (Pearson)"
)
```

```
ggcorrplot::corrplot(Num.cor_2, method = "number", type = "lower")



# SPLIT ----------------------------------------------------------------

rm(list = ls())

gc()

Data_SinCorr <- read.xlsx("G://Proyectos/Modelo de Retencion/SALIDA/Data_SinCrr.xlsx")

Data_SinCorr <- Data_SinCorr %>%

 select(-FECHA_INICIO, -FECHA_FIN, -CHASIS, -FEC_NACIMIENTO_ASEG, -CLASE_SPTO, -IMP_PRIMA_ANUL,

       -FEC_ANULACION, -ALTURA_ANUL, -VA_NUEVO, -NUM_POLIZA, -ID, -FASECOLDA, -PLACA, -IMP_DESC_COMERCIAL, -PRIMA_PRERENOV,

       -MCA_VIG, -MCA_ANUL, -FECHA_EMISION, -SUAVIZACION) %>%

  mutate(RC = if_else(is.na(RC), 1000, RC))

split1 <- sort(sample(nrow(Data_SinCorr), nrow(Data_SinCorr)*0.7))

train <- Data_SinCorr[split1,]

test <- Data_SinCorr[-split1,]


summary(Data_SinCorr)



# Script Modelo ------------------------------------------------------------


gc()
```

```
# 1. GLM ----------------------------------------------------------------



str(Data_SinCorr)

summary(Data_SinCorr)


Logit <- glm(formula = MCA_RENOV ~ TIPO_VEHICULO + SINIESTROS + BONUS_MALUS + FEC_RENOV + NEGOCIO + MARCA +

          DEPARTAMENTO + RC + FIDELIZACION + ANTIGUEDAD + VAR_PRIMA +

        RANGO_ANTIGUEDAD + REGIONAL + ALTURA_RENOV + RANGO_PRIMA_ANT + RANGO_SCORE + GENERO + RANGO_VAR_PRIMA ,

        data = Data_SinCorr,

        family = 'binomial')


summary(Logit)


Predict <- predict(Logit, Data_SinCorr, type = 'response')


table <- data.frame(cut.off = double(),

                Accurancy = numeric(),

                Sensitivity = numeric(),

                Specificity = numeric(),

                Kappa = numeric(),

                stringsAsFactors = FALSE)


co <- seq(0, 1, 0.05)


for (i in 1:length(co)) {
```

```
  predict_labels <- as.factor(as.character(ifelse(Predict > co[i], 1,0)))
  confumat <- confusionMatrix(data = predict_labels,
                              reference = as.factor(Data_SinCorr$MCA_RENOV))

  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
}

table <- table %>%
  arrange(desc(Kappa*Specificity))

best_cutoff <- table[1, 1]

Data_SinCorr$MCA_RENOV <- as.factor(Data_SinCorr$MCA_RENOV)
Predict <- as.factor(ifelse(Predict > best_cutoff, 1, 0))

confusionMatrix(Data_SinCorr$MCA_RENOV, Predict)
```

```
   ##quitando algunas variables con datos de signifcacia similares y menor cantidad de variables
Logit <- glm(formula = MCA_RENOV ~ TIPO_VEHICULO + BONUS_MALUS + ALTURA_RENOV + RANGO_PRIMA_ANT + GENERO ,
          data = Data_SinCorr,
```

```
            family = 'binomial')

summary(Logit)

Predict <- predict(Logit, Data_SinCorr, type = 'response')

table <- data.frame(cut.off = double(),
                    Accurancy = numeric(),
                    Sensitivity = numeric(),
                    Specificity = numeric(),
                    Kappa = numeric(),
                    stringsAsFactors = FALSE)

co <- seq(0, 1, 0.05)

for (i in 1:length(co)) {
  predict_labels <- as.factor(as.character(ifelse(Predict > co[i], 1,0)))
  confumat <- confusionMatrix(data = predict_labels,
                          reference = as.factor(Data_SinCorr$MCA_RENOV))

  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
}
```

```
table <- table %>%
  arrange(desc(Kappa*Specificity))

best_cutoff <- table[1, 1]

Data_SinCorr$MCA_RENOV <- as.factor(Data_SinCorr$MCA_RENOV)
Predict <- as.factor(ifelse(Predict > best_cutoff, 1, 0))

confusionMatrix(Data_SinCorr$MCA_RENOV, Predict)

## Prueba de pesos
Weight_Train_list <- list()
Weight_Test_list <- list()

w1 <- seq(0.05, 1, 0.05)
w0 <- 1-w1
w0[length(w0)] <- 1

w <- rep(NA, length(train))

for (j in 1:length(w0)) {
  w[which(train$MCA_RENOV == 1)] <- w1[j]
  w[which(train$MCA_RENOV == 0)] <- w0[j]

  summary(w)

  str(train)
```

```
summary(train)

Logit <- glm(formula = MCA_RENOV ~ TIPO_VEHICULO + BONUS_MALUS + ALTURA_RENOV + RANGO_PRIMA_ANT + GENERO ,
             data = train,
             family = 'binomial')

summary(Logit)

Predict <- predict(Logit, test, type = 'response')

table <- data.frame(cut.off = double(),
                    Accurancy = numeric(),
                    Sensitivity = numeric(),
                    Specificity = numeric(),
                    Kappa = numeric(),
                    stringsAsFactors = FALSE)

co <- seq(0, 1, 0.05)

for (i in 1:length(co)) {
  predict_labels <- as.factor(as.character(ifelse(Predict > co[i], 1,0)))
  confumat <- confusionMatrix(data = predict_labels,
                              reference = as.factor(test$MCA_RENOV))

  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
```

```
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
 }

 table <- table %>%
  arrange(desc(Kappa*Specificity))

 best_cutoff <- table[1, 1]


 Predict_train <- as.factor(ifelse(Logit$fitted.values > best_cutoff, 1, 0))
 Predict_test <- as.factor(ifelse(Predict > best_cutoff, 1, 0))

 train$MCA_RENOV <- as.factor(train$MCA_RENOV)
 test$MCA_RENOV <- as.factor(test$MCA_RENOV)

 Weight_Train <- confusionMatrix(train$MCA_RENOV, Predict_train)
 Weight_Test <- confusionMatrix(test$MCA_RENOV, Predict_test)

 print(paste0('Calculo ', j, ' de ', length(w0), ' realizado correctamente'))


}



  ##Modelo elegido de glm
```

```
str(train)
summary(train)

test <- test %>%
  filter(FEC_RENOV >= 202208)

Logit <- glm(formula = MCA_RENOV ~ TIPO_VEHICULO + BONUS_MALUS + ALTURA_RENOV + RANGO_PRIMA_ANT + GENERO ,
             data = train,
             family = 'binomial')

summary(Logit)

Predict <- predict(Logit, test, type = 'response')

table <- data.frame(cut.off = double(),
                    Accurancy = numeric(),
                    Sensitivity = numeric(),
                    Specificity = numeric(),
                    Kappa = numeric(),
                    stringsAsFactors = FALSE)

co <- seq(0, 1, 0.05)

for (i in 1:length(co)) {
  predict_labels <- as.factor(as.character(ifelse(Predict > co[i], 1,0)))
  confumat <- confusionMatrix(data = predict_labels,
                              reference = as.factor(test$MCA_RENOV))
```

```
  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
}


table <- table %>%
  arrange(desc(Kappa*Specificity))

best_cutoff <- table[1, 1]


Predict_train <- as.factor(ifelse(Logit$fitted.values > best_cutoff, 1, 0))
Predict_test <- as.factor(ifelse(Predict > best_cutoff, 1, 0))

train$MCA_RENOV <- as.factor(train$MCA_RENOV)
test$MCA_RENOV <- as.factor(test$MCA_RENOV)


summary(Logit)
confusionMatrix(train$MCA_RENOV, Predict_train)
confusionMatrix(test$MCA_RENOV, Predict_test)

   ## Cross Validation
```

```
Data_SinCorr$MCA_RENOV <- as.integer(as.character(Data_SinCorr$MCA_RENOV))

Data_Result <- Data_SinCorr

Data_Result$pred <- NA

k <- 10

set.seed(9)

Index <- sample(k, nrow(Data_SinCorr), replace = T, prob = rep(1/k, k))

for (i in 1:k) {

  trainData <- Data_SinCorr[- (Index == i), ]

  testData <- Data_SinCorr[Index == i, ]



  w <- rep(NA, length(trainData))

  w[which(trainData$MCA_RENOV == 1)] <- 0.05

  w[which(trainData$MCA_RENOV == 0)] <- 0.05

  summary(w)

  Logit <- glm(formula = MCA_RENOV ~ TIPO_VEHICULO + BONUS_MALUS + ALTURA_RENOV + RANGO_PRIMA_ANT + GENERO,

               data = trainData,

               family = 'binomial',

               weights = w)
```

```
  pred <- predict(Logit, testData, type = 'response')

  Data_Result[Index == i, ]$pred <- pred

}


  table <- data.frame(cut.off = double(),
                      Accuracy = numeric(),
                      Sensitivity = numeric(),
                      Specificity = numeric(),
                      Kappa = numeric(),
                      stringsAsFactors = FALSE)

  co <- seq(0.05, 1, 0.05)

  for (i in 1:length(co)) {

    predict_labels <- as.factor(as.character(ifelse(Data_Result$pred > co[i], 1, 0)))

    confumat <- confusionMatrix(data = predict_labels,
                                reference = as.factor(Data_Result$MCA_RENOV))

    table[i,1] <- co[i]
    table[i,2] <- round(confumat$overall[1]*100, 2)
    table[i,3] <- round(confumat$byClass[1]*100, 2)
    table[i,4] <- round(confumat$byClass[2]*100, 2)
```

```
  table[i,5] <- round(confumat$overall[2]*100, 2)
}

table <- table %>%
  arrange(desc(Kappa*Specificity))

best_cutoff <- table[1, 1]

Data_Result$pred_def <- as.factor(ifelse(Data_Result$pred > best_cutoff, 1, 0))
Data_Result$pred_def <- as.factor(as.character(Data_Result$pred_def))
Data_Result$MCA_RENOV <- as.factor(Data_Result$MCA_RENOV)

Data_Result <- Data_Result %>%
  filter(FEC_RENOV >= 202210)

MatrizConf <- confusionMatrix(Data_Result$MCA_RENOV, Data_Result$pred_def)
MatrizConf

## Análisis por variable de renovación sobre la base de test

Data_Result <- Data_Result %>%
  arrange((pred)) %>%
  mutate(n = ceiling(c(1:31713)/(3172/2))) %>%
```

```
  mutate(MCA_RENOV = as.integer(MCA_RENOV))


 Data_Result_val <- Data_Result %>% group_by(n) %>%  summarise(pred = mean(pred), real = mean(MCA_RENOV-1), count = n())


 plot(x = Data_Result_val$n,
     y = Data_Result_val$real,
     xlab ="Buckets por Probabilidad de Renovación",
     ylab = "Probabilidad de Renovación",
     main = "Lift Curve por buckets de Renovación en GLM",
     pch = 19)



 ## RESULTADOS POR VARIABLE


 ## TIPO VEHICULO
 TABLE1 <- Data_Result %>% group_by(TIPO_VEHICULO) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE1, aes(x = TIPO_VEHICULO,
                    y = Probabilidad,
                    fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
```

```
        position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Tipo de Vehiculo")


 ## BONUS MALUS
 TABLE2 <- Data_Result %>% group_by(BONUS_MALUS) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE2, aes(x = BONUS_MALUS,
                    y = Probabilidad,
                    fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
        position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Bonus Malus")


 ## ALTURA RENOVACION
 TABLE3 <- Data_Result %>% group_by(ALTURA_RENOV) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE3, aes(x = ALTURA_RENOV,
                    y = Probabilidad,
                    color = Cat)) +
```

```
  geom_line(stat = "identity",
            linetype = 1,
            lwd = 1.1) +
  ylim(c(0.7,1)) +
  ggtitle("Resultado Modelo GLM por Altura del Riesgo")


 ## RANGO PPRIMA ANTERIOR
 TABLE4 <- Data_Result %>% group_by(RANGO_PRIMA_ANT) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE4, aes(x = RANGO_PRIMA_ANT,
                    y = Probabilidad,
                    fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Rango de Prima")


 ##GENERO
 TABLE5 <- Data_Result %>% group_by(GENERO) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')
```

```
ggplot(TABLE5, aes(x = GENERO,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Genero")


## FEC_RENOV
TABLE6 <- Data_Result %>% group_by(FEC_RENOV) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE6, aes(x = FEC_RENOV,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Fecha de Renovación")


## DEPARTAMENTO
TABLE7 <- Data_Result %>% group_by(DEPARTAMENTO) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1), n = n()) %>%
```

```
  arrange(desc(n)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad') %>%
  select(-n) %>%
  head(20)


ggplot(TABLE7, aes(x = DEPARTAMENTO,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Fecha de Departamento")


## MARCA
TABLE8 <- Data_Result %>% group_by(MARCA) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1), n = n()) %>%
  arrange(desc(n)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad') %>%
  select(-n) %>%
  head(20)


ggplot(TABLE8, aes(x = MARCA,
                   y = Probabilidad,
```

```
                fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Fecha de Marca")


## ANTIGUEDAD
TABLE9 <- Data_Result %>% group_by(ANTIGUEDAD) %>%  summarise(Prediccion
= mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to =
'Probabilidad') %>%
  head(60)


ggplot(TABLE9, aes(x = ANTIGUEDAD,
                y = Probabilidad,
                color = Cat)) +
  geom_line(stat = "identity",
            linetype = 1,
            lwd = 1.1) +
  ylim(c(0.7,1)) +
  ggtitle("Resultado Modelo GLM por Antiguedad")


## FIDELIZACION
TABLE10 <- Data_Result %>% group_by(FIDELIZACION) %>% summarise(Prediccion
= mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
```

```
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE10, aes(x = FIDELIZACION,
                     y = Probabilidad,
                     fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Fidelización")


 ## REGIONAL
 TABLE11 <- Data_Result %>% group_by(REGIONAL) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


 ggplot(TABLE11, aes(x = REGIONAL,
                     y = Probabilidad,
                     fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Fecha de Regional")
```

```
## RANGO VAR PRIMA

TABLE12 <- Data_Result %>% group_by(RANGO_VAR_PRIMA) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

  ungroup() %>%

  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE12, aes(x = RANGO_VAR_PRIMA,

                    y = Probabilidad,

                    fill = Cat)) +

  geom_bar(stat = "identity", position = position_dodge())+

  ylim(c(0,1)) +

  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",

            position = position_dodge(0.9), size=4)+

  ggtitle("Resultado Modelo GLM por Fecha de Rango de Variación de Prima")


## NEGOCIO

TABLE13 <- Data_Result %>% group_by(NEGOCIO) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

  ungroup() %>%

  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE13, aes(x = NEGOCIO,

                    y = Probabilidad,

                    fill = Cat)) +

  geom_bar(stat = "identity", position = position_dodge())+

  ylim(c(0,1)) +

  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
```

```
          position = position_dodge(0.9), size=4)+
   ggtitle("Resultado Modelo GLM por Fecha de Tipo de Negocio")



# 2. Random Forest ------------------------------------------------------
gc()
  Data_SinCorr$MCA_RENOV = as.factor(Data_SinCorr$MCA_RENOV)
RF <- randomForest(MCA_RENOV ~ TIPO_VEHICULO + SINIESTROS + BONUS_MALUS
+ FEC_RENOV + NEGOCIO + MARCA +
                     DEPARTAMENTO + RC + FIDELIZACION + ANTIGUEDAD + VAR_PRIMA +
                     RANGO_ANTIGUEDAD + REGIONAL + ALTURA_RENOV +
RANGO_PRIMA_ANT +  GENERO + RANGO_VAR_PRIMA,
                   data = Data_SinCorr,
                   ntree = 100,
                   importance = TRUE)

Predict <- predict(RF, Data_SinCorr, 'prob')

table <- data.frame(cut.off = double(),
                     Accurancy = numeric(),
                     Sensitivity = numeric(),
                     Specificity = numeric(),
                     Kappa = numeric(),
                     stringsAsFactors = FALSE)

co <- seq(0, 1, 0.05)
```

```
for (i in 1:length(co)) {
  predict_labels <- as.factor(as.character(ifelse(Predict[,2] > co[i], 1, 0)))

  confumat <- confusionMatrix(data = predict_labels,
                              reference = Data_SinCorr$MCA_RENOV)

  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
}

table <- table %>%
  arrange(desc(Kappa * Specificity))

best_cutoff <- table[1, 1]

Predict <- as.factor(ifelse(Predict[,2] > best_cutoff, 1, 0))

confusionMatrix(Predict, Data_SinCorr$MCA_RENOV)

RF$importance

Plot_importance <- data.frame(RF$importance) %>%
  mutate(Variable = as.factor(row.names(RF$importance))) %>%
  arrange(desc(MeanDecreaseAccuracy)) %>%
```

```
  select(Variable, MeanDecreaseAccuracy, MeanDecreaseGini)


ggplot(data = Plot_importance,
       aes(x = MeanDecreaseAccuracy,
           y = reorder(Variable, MeanDecreaseAccuracy),
           fill = MeanDecreaseAccuracy)) +
                geom_col() +
                ggtitle('Importancia Random Forest') +
                scale_fill_continuous(guide='none') +
         labs(x= 'Mean Decrease Accuracy', y = 'Variables')+
         theme_bw() +
  theme(plot.title = element_text(size = 20))


     ## Se retiran algunas variables con el fin de generar signifcancias similar usando
menos vairbles


gc()
Data_SinCorr$MCA_RENOV = as.factor(Data_SinCorr$MCA_RENOV)



RF <- randomForest(MCA_RENOV ~    BONUS_MALUS + FEC_RENOV + NEGOCIO +
MARCA +
                    DEPARTAMENTO +  FIDELIZACION + ANTIGUEDAD + VAR_PRIMA +
                   REGIONAL + ALTURA_RENOV + RANGO_PRIMA_ANT +  GENERO ,
                  data = Data_SinCorr,
                  ntree = 100,
                  mtry = 3,
                  nodesize = 4,
```

```
                importance = TRUE)

Predict <- predict(RF, Data_SinCorr, 'prob')

table <- data.frame(cut.off = double(),
                    Accurancy = numeric(),
                    Sensitivity = numeric(),
                    Specificity = numeric(),
                    Kappa = numeric(),
                    stringsAsFactors = FALSE)

co <- seq(0, 1, 0.05)

for (i in 1:length(co)) {
  predict_labels <- as.factor(as.character(ifelse(Predict[,2] > co[i], 1, 0)))

  confumat <- confusionMatrix(data = predict_labels,
                          reference = Data_SinCorr$MCA_RENOV)

  table[i,1] <- co[i]
  table[i,2] <- round(confumat$overall[1]*100, 2)
  table[i,3] <- round(confumat$byClass[1]*100, 2)
  table[i,4] <- round(confumat$byClass[2]*100, 2)
  table[i,5] <- round(confumat$overall[2]*100, 2)
}

table <- table %>%
```

```
  arrange(desc(Kappa * Specificity))

best_cutoff <- table[1, 1]

Predict <- as.factor(ifelse(Predict[,2] > best_cutoff, 1, 0))

confusionMatrix(Predict, Data_SinCorr$MCA_RENOV)

RF$importance

Plot_importance <- data.frame(RF$importance) %>%
  mutate(Variable = as.factor(row.names(RF$importance))) %>%
  arrange(desc(MeanDecreaseAccuracy)) %>%
  select(Variable, MeanDecreaseAccuracy, MeanDecreaseGini)

ggplot(data = Plot_importance,
       aes(x = MeanDecreaseAccuracy,
           y = reorder(Variable, MeanDecreaseAccuracy),
           fill = MeanDecreaseAccuracy)) +
  geom_col() +
  ggtitle('Importancia Random Forest') +
  scale_fill_continuous(guide='none') +
  labs(x= 'Mean Decrease Accuracy', y = 'Variables')+
  theme_bw() +
  theme(plot.title = element_text(size = 20))
```

```
   ## Cross validation del random forest

Data_SinCorr <- Data_SinCorr %>%
  select(MCA_RENOV , BONUS_MALUS , FEC_RENOV , NEGOCIO , MARCA ,
         DEPARTAMENTO ,  FIDELIZACION , ANTIGUEDAD , VAR_PRIMA ,
         REGIONAL , ALTURA_RENOV , RANGO_PRIMA_ANT ,  GENERO)


train <- train %>%
  select(MCA_RENOV , BONUS_MALUS , FEC_RENOV , NEGOCIO , MARCA ,
         DEPARTAMENTO ,  FIDELIZACION , ANTIGUEDAD , VAR_PRIMA ,
         REGIONAL , ALTURA_RENOV , RANGO_PRIMA_ANT ,  GENERO)

test <- test %>%
  select(MCA_RENOV , BONUS_MALUS , FEC_RENOV , NEGOCIO , MARCA ,
         DEPARTAMENTO ,  FIDELIZACION , ANTIGUEDAD , VAR_PRIMA ,
         REGIONAL , ALTURA_RENOV , RANGO_PRIMA_ANT ,  GENERO)


Data_Result <- Data_SinCorr
Data_Result$pred <- NA

k <- 10

set.seed(9)
Index <- sample(k, nrow(Data_SinCorr), replace = T, prob = rep(1/k, k))
```

```
for (i in 1:k) {
  testData <- Data_SinCorr[Index == i, ]
  trainData <- Data_SinCorr[-(Index == i), ]

  rf <- randomForest(x = trainData[, -1],
                     y = trainData[, 1],
                     ntree = 100,
                     mtry = 3,
                     nodesize = 4)
  pred <- predict(rf, testData, 'prob')

  Data_Result[Index == i, ]$pred <- pred[, 2]

  print(paste0("Ok Cross Validation. ejecución al ", i*100/k, " %"))

}

table <- data.frame(cut.off = double(),
                    Accuracy = numeric(),
                    Sensitivity = numeric(),
                    Specificity = numeric(),
                    Kappa = numeric(),
                    stringsAsFactors = FALSE)

co <- seq(0.5, 1, 0.5)

for (i in 1:length(co)) {
```

```
  predict_labels <- as.factor(as.character(ifelse(Data_Result$pred > co[i], 1, 0)))

  confumat <- confusionMatrix(data = predict_labels,
                              reference = Data_SinCorr$MCA_RENOV)

  table[i, 1] <- co[i]
  table[i, 2] <- round(confumat$overall[1]*100, 2)
  table[i, 3] <- round(confumat$byClass[1]*100, 2)
  table[i, 4] <- round(confumat$byClass[2]*100, 2)
  table[i, 5] <- round(confumat$overall[2]*100, 2)

}

table <- table %>%
  arrange(desc(Kappa)); table

best_cutoff <- table[1, 1]

Data_Result <- Data_Result %>%
  filter(FEC_RENOV >= 202210)

Data_Result$pred_def <- ifelse(Data_Result$pred > best_cutoff, 1, 0)
Data_Result$pred_def <- as.factor(as.character(Data_Result$pred_def))

confumat <- confusionMatrix(data = Data_Result$pred_def,
                            reference = Data_Result[, 1])
confumat
```

```
## Análisis por variable de renovación sobre la base de test


Data_Result <- Data_Result %>%

  arrange((pred)) %>%

  mutate(n = ceiling(c(1:31713)/(3172/2))) %>%

  mutate(MCA_RENOV = as.integer(MCA_RENOV))


Data_Result_val <- Data_Result %>% group_by(n) %>%  summarise(pred = mean(pred), real = mean(MCA_RENOV-1), count = n())


plot(x = Data_Result_val$n,

    y = Data_Result_val$real,

    xlab ="Buckets por Probabilidad de Renovación",

    ylab = "Probabilidad de Renovación",

    main = "Lift Curve por buckets de Renovación en Random Forest",

    pch = 19)



## RESULTADOS POR VARIABLE




## BONUS MALUS

TABLE2 <- Data_Result %>% group_by(BONUS_MALUS) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

  ungroup() %>%
```

```
 pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE2, aes(x = BONUS_MALUS,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Bonus Malus")


## ALTURA RENOVACION
TABLE3 <- Data_Result %>% group_by(ALTURA_RENOV) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE3, aes(x = ALTURA_RENOV,
                   y = Probabilidad,
                   color = Cat)) +
  geom_line(stat = "identity",
            linetype = 1,
            lwd = 1.1) +
  ylim(c(0.7,1)) +
  ggtitle("Resultado Modelo Random Forest por Altura del Riesgo")
```

```
## RANGO PPRIMA ANTERIOR

TABLE4 <- Data_Result %>% group_by(RANGO_PRIMA_ANT) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

 ungroup() %>%

 pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE4, aes(x = RANGO_PRIMA_ANT,

                   y = Probabilidad,

                   fill = Cat)) +

 geom_bar(stat = "identity", position = position_dodge())+

 ylim(c(0,1)) +

 geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",

           position = position_dodge(0.9), size=4)+

 ggtitle("Resultado Modelo Random Forest por Rango de Prima")


##GENERO

TABLE5 <- Data_Result %>% group_by(GENERO) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

 ungroup() %>%

 pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE5, aes(x = GENERO,

                   y = Probabilidad,

                   fill = Cat)) +

 geom_bar(stat = "identity", position = position_dodge())+

 ylim(c(0,1)) +

 geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
```

```
        position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo GLM por Genero")


## FEC_RENOV
TABLE6 <- Data_Result %>% group_by(FEC_RENOV) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE6, aes(x = FEC_RENOV,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
        position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fecha de Renovación")


## DEPARTAMENTO
TABLE7 <- Data_Result %>% group_by(DEPARTAMENTO) %>% summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1), n = n()) %>%
  arrange(desc(n)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad') %>%
  select(-n) %>%
  head(20)
```

```
ggplot(TABLE7, aes(x = DEPARTAMENTO,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fecha de Departamento")


## MARCA
TABLE8 <- Data_Result %>% group_by(MARCA) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1), n = n()) %>%
  arrange(desc(n)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad') %>%
  select(-n) %>%
  head(20)


ggplot(TABLE8, aes(x = MARCA,
                   y = Probabilidad,
                   fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fecha de Marca")
```

```
## ANTIGUEDAD

TABLE9 <- Data_Result %>% group_by(ANTIGUEDAD) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

 ungroup() %>%

 pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad') %>%

 head(60)


ggplot(TABLE9, aes(x = ANTIGUEDAD,

                   y = Probabilidad,

                   color = Cat)) +

 geom_line(stat = "identity",

           linetype = 1,

           lwd = 1.1) +

 ylim(c(0.7,1)) +

 ggtitle("Resultado Modelo Random Forest por Antiguedad")


## FIDELIZACION

TABLE10 <- Data_Result %>% group_by(FIDELIZACION) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%

 ungroup() %>%

 pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE10, aes(x = FIDELIZACION,

                    y = Probabilidad,

                    fill = Cat)) +

 geom_bar(stat = "identity", position = position_dodge())+

 ylim(c(0,1)) +
```

```
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fidelización")


## REGIONAL
TABLE11 <- Data_Result %>% group_by(REGIONAL) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')


ggplot(TABLE11, aes(x = REGIONAL,
                    y = Probabilidad,
                    fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fecha de Regional")



## NEGOCIO
TABLE13 <- Data_Result %>% group_by(NEGOCIO) %>%  summarise(Prediccion = mean(pred), Real = mean(MCA_RENOV-1)) %>%
  ungroup() %>%
  pivot_longer(cols = c('Prediccion', 'Real'), names_to = 'Cat', values_to = 'Probabilidad')
```

```
ggplot(TABLE13, aes(x = NEGOCIO,
                    y = Probabilidad,
                    fill = Cat)) +
  geom_bar(stat = "identity", position = position_dodge())+
  ylim(c(0,1)) +
  geom_text(aes(label= round(Probabilidad,2)), vjust=-0.5, color="black",
            position = position_dodge(0.9), size=4)+
  ggtitle("Resultado Modelo Random Forest por Fecha de Tipo de Negocio")
```