UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

# DESARROLLO DE UN MODELO PREDICTIVO DE ASIGNACION DE COBRANZA JUDICIAL APLICADO A UNA EMPRESA DE CREDITO AUTOMOTRIZ.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

EDUARDO IGNACIO ILABACA MEZA

PROFESORA GUÍA:

ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:

CAROLINA SEGOVIA RIQUELME

FELIPE VILDOSO CASTILLO

SANTIAGO DE CHILE

2023

**RESUMEN DE LA MEMORIA PARA OPTAR AL**

**TÍTULO DE:** Ingeniero Civil Industrial.

**POR:** Eduardo Ignacio Ilabaca Meza

**FECHA:** 2023

**PROFESORA GUÍA:** Alejandra Puente Chandía

# DESARROLLO DE UN MODELO PREDICTIVO DE ASIGNACION DE COBRANZA JUDICIAL APLICADO A UNA EMPRESA DE CREDITO AUTOMOTRIZ.

El trabajo se realiza en una empresa de crédito automotriz, la cual ofrece productos de financiación de automotores tanto a personas naturales como a concesionarios. Uno de sus productos, consiste en un contrato de financiamiento en cuotas para adquirir un vehículo, dentro de este contrato, existe la posibilidad de que las entidades subscriptas presenten atrasos o cesen el pago de estas cuotas pactadas, de ser este el caso, es posible aplicar 2 métodos distintos para realizar la cobranza de estas cuotas, la asignación sobre que método se aplicara a cada caso depende de parámetros definidos.

El objetivo del trabajo de memoria consiste en mejorar la asignación de los métodos de cobranza, mediante el uso de modelos predictivos para determinar la mejor estrategia para cada cliente, con el fin de mejorar el uso de recursos y las probabilidades de recuperación de deuda. Para llevar a cabo el trabajo se utiliza la metodología CRISP-DM, en la cual se prueban distintos modelos de *machine learning* para cumplir el objetivo propuesto.

Se realizan un total de 6 modelos, se consideran variables tanto de origen del contrato como variables de comportamiento de los usuarios. Se selecciona el modelo que presenta la mejor capacidad predictiva, correspondiente a un modelo de regresión lineal (GLM). Los resultados de este modelo presentan un 64% de precisión en los resultados sobre predecir si un cliente presentara pagos o no a los 3 meses en la etapa de cobranza. El AUC del modelo es de un 65%, lo cual es superior a un modelo de asignación aleatorio, pero presenta oportunidades de mejora a la hora de generar una predicción más robusta. Las variables relevantes muestran características del cliente hacia el término del contrato, siendo en particular muy importantes aquellas que capturan el comportamiento del cliente los últimos 3 meses antes de ser castigado. En conclusión, el modelo presenta mejoras frente a la situación actual, ya que posee mayor capacidad de predicción.

Como trabajo futuro se propone capturar y generar más variables anteriores a la fecha de castigo, en particular sobre los últimos 3 meses y agregar variables externas relacionadas a deudas o variables macroeconómicas con el fin de aumentar el desempeño general del modelo y, finalmente, realizar un seguimiento a los recuperos monetarios de los grupos asignados en base al modelo predictivo.

# DEDICATORIA

*A mi Madre,*

# AGRADECIMIENTOS

*Quiero agradecer a todos quienes me acompañaron estos años,*

*A mi Familia que estuvo siempre presente,*

*A mis primos,*

*A mis amigos,*

*A Diego e Ignacio,*

*Gracias por estar en los momentos buenos y los no tan buenos,*

*Sin ustedes este camino hubiera sido muy diferente.*

# TABLA DE CONTENIDO

# INDICE DE FIGURAS

# INDICE DE TABLAS

# 1. ANTECEDENTES GENERALES

## 1.1. Antecedentes de la Empresa e Industria

### 1.1.1. Identificación del y descripción del sector industrial

General Motors Financial se desempeña en el sector industrial correspondiente al financiamiento del sector automotriz de Chile, esta industria se compone principalmente de agrupaciones nacionales, las cuales poseen la representación de una o más marcas de vehículos en el país, incluyendo la comercialización tanto de vehículos nuevos como usados. En el pasado año 2021, las ventas de automóviles livianos y medianos alcanzaron un nivel de 415.581 unidades, representando un aumento del 60,6% con respecto al año anterior, siendo esto explicado principalmente por una mayor liquidez en el mercado chileno.

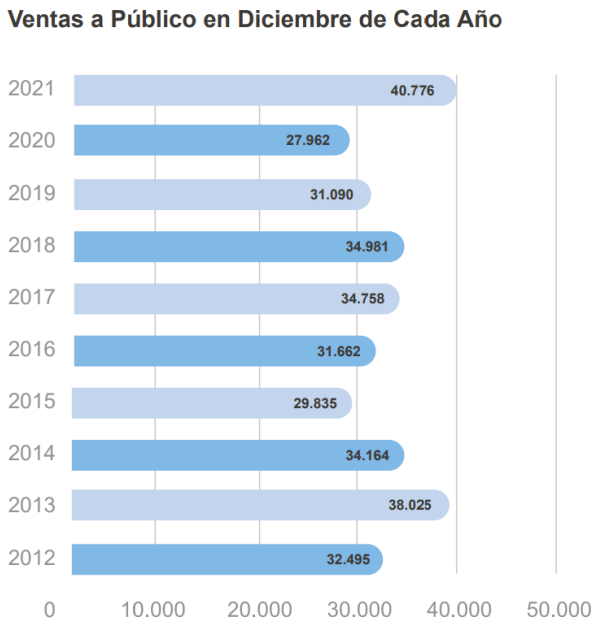



*Figura 1: Mercado automotor. ANAC diciembre 2021*

En este mismo año, GM Financial registró un total de 17.100 unidades financiadas. Si se toma en cuenta el giro de negocio de GM Financial, el cual es, el financiamiento solo de la marca Chevrolet, se considera que la participación de mercado fue de un 42,77%. Los principales competidores de GMF para el negocio de financiamiento automotriz de la red de concesionarios GM son:

- Forum Servicios Financieros
- Santander Consumer
- Global Soluciones Financieras
- Cajas de Compensación

### 1.1.2. Tipo de Organización

General Motors Financial es una Sociedad Anónima Cerrada. El capital de la compañía se encuentra divida en 4000 acciones, con una distribución de 99,975% perteneciente a General Motors Financial Company Inc., y un 0,025% perteneciente a GMF International LLC, a su vez, General Motors Financial Company Inc., es poseída en un 100% por GM Corporation. En cuanto al gobierno corporativo, el directorio al 31 de diciembre de 2021 es presidido por Paul Holtgreive, compuesto además por 7 miembros entre directores y directores suplentes. Se declaran también a la misma fecha un total de 132 empleados, los cuales se dividen en 22 personas correspondientes a personal de gerencia y 110 personas entre supervisores y analistas.

GM Financial presta servicios de financiamiento tanto a clientes particulares (personas naturales) como a concesionarios de la marca Chevrolet. La empresa cuenta con 19 concesionarios a lo largo de Chile, en particular 12 dedicado a la venta de automóviles, quienes a su vez cuentan con 69 puntos de venta de la marca Chevrolet, considerando también 7 puntos de venta dedicados a la distribución de repuestos.

En el aspecto financiero, General Motors Financial Company INC. (Compañía global), reportó:

- Ingresos antes de impuestos por USD 5.036 millones.
- Originación de cartera de crédito y leasing por USD 50.889 millones.
- Patrimonio de USD 14.387 millones
- Liquidez por USD 26.749 millones.
- Activos productivos por USD 102.794 millones.

### 1.1.3. Servicios y procesos

Como se mencionó anteriormente, GM Financial se especializa en el financiamiento automotriz, ofreciendo créditos y leasing a personas naturales y empresas para la adquisición de vehículos livianos y pesados, actualmente este servicio de financiamiento se puede dividir en dos grandes grupos de líneas de negocio, los cuales son:

**Plan Mayor o Wholesale**

Plan enfocado a otorgar financiamiento a concesionarios, GM Financial ha ofrecido consistentemente servicios por 41 años a los diferentes concesionarios de General Motors en Chile. A la fecha se cuenta con 19 concesionarios en el país de la marca Chevrolet, quienes a su vez cuentan con 69 puntos de venta. La compañía mantiene una profunda relación de negocios con los concesionarios, con el fin de desarrollar la venta de automóviles Chevrolet en el país.

**Plan Menor o Retail**

Enfocado a los clientes finales (personas naturales), el Plan Menor o Retail, ofrece un amplio portafolio de opciones de financiamiento, buscando incrementar los márgenes de utilidad de todos los concesionarios de General Motors. El negocio se basa en la originación de créditos bajo los lineamientos corporativos de riesgo. A cierre del año 2021 GM Financial cuenta con más de 48.000 clientes a lo largo del país, generando un portafolio de activos muy atomizado, reduciendo así los riesgos de concentración de un tipo particular de cliente. Mediante esta línea de negocia se financiaron 17.339 unidades nuevas sobre el universo de 415.581 unidades de autos vendidas en todo Chile.

### 1.1.4. Clientes y usuarios

El cliente de GM Financial será toda persona o empresa que origine un contrato de financiamiento para vehículos o flotas mediante Chevrolet Servicios Financieros.

## 1.2. Descripción del problema y justificación

Un cliente durante el transcurso de su contrato puede presentar morosidad al momento de pagar sus cuotas. Dependiendo de cuantos días de morosidad posea un contrato se definen los siguientes *buckets*[1]:


*Figura 2: Buckets de morosidad.*

El número que acompaña a cada Bucket representa que intervalo en días de morosidad considera, por ejemplo, el Bucket 0-30 contiene todas las cuentas que presenten de 0 a 30 días de morosidad, el Bucket 31-60 contiene aquellas cuentas que poseen de 31 a 60 días de morosidad, etc. En cada uno de estos Buckets se utilizan distintas estrategias de retención, las cuales pueden ser, repactar la cuota, extensión de contrato, refinanciamiento, entre otras estrategias que dependerán de la severidad de la mora. Sin embargo, cuando las cuentas ya poseen más de 120 días de morosidad, son traspasadas al área de *Salvage*. En esta etapa, la cuenta es castigada, significando esto, que se vuelven exigibles financiera y legalmente todas las cuotas restantes hacia el fin del contrato como un monto total.

Dentro del área de Salvage existen actualmente dos estrategias distintas para la gestión de las cuentas, siendo estas:

### 1.2.1. Proceso regular

Las cuentas en este proceso empiezan sus gestiones en agencias de cobranza externas, en donde se buscan acuerdos de pago con el cliente, si estas agencias no tienen éxito

[1] Grupos de cuentas en un rango de morosidad determinado.

en 80 días, las cuentas son traspasadas a estudios de abogados, en esta etapa es tramitada la notificación al cliente de que está enfrentando un proceso legal, en donde se realizara una reposición del vehículo de manera voluntaria o involuntaria.

### 1.2.2. Proceso Demanda anticipada

Las cuentas destinadas para este proceso empiezan sus gestiones directamente en las agencias de abogados, siguiendo también con los pasos de notificación y reposición del vehículo. Estas cuentas no pasaran los 80 días en agencias como el proceso anterior.

La diferencia de los procesos se puede observar en la figura 3:

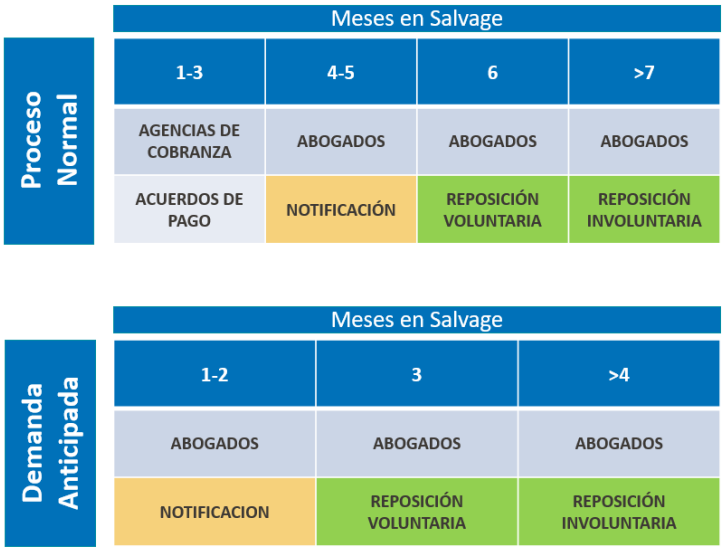



*Figura 3: Estrategias Salvage. Etapas esperadas*

En un origen, todas las cuentas en Salvage estaban destinadas al proceso normal, sin embargo, en agosto de 2020, se implementa un modelo *PCA*[2], el cual determina cada mes que cuentas son designadas a las estrategias mencionadas anteriormente. Hoy en día, no se tiene información certera sobre el real impacto de esta segmentación, puesto que no se establecieron grupos de control definidos, sumado a que el despliegue de este modelo coincidió con el principio de los retiros monetarios de las AFP (Administradoras de Fondos de Pensión), provocando una mayor liquidez en el mercado. Desde este hito las cantidades de perdidas monetarias disminuyeron y los saldos recuperados incrementaron notablemente, haciendo este periodo incomparable con otros en la línea de tiempo. Durante este año, este modelo ha designado en promedio un 15% de las cuentas que llegan a Salvage directamente al Proceso de demanda anticipada y un 85% al Proceso Regular.

[2] Principal Component Analysis: Técnica utilizada para describir un conjunto de datos en términos de variables no correlacionadas, con el fin de encontrar un menor número de variables que expliquen aproximadamente en la misma proporción un fenómeno en la misma medida que sus variables originales.

Tomando en cuenta cuentas desde el año 2018 hasta diciembre 2021, el global resultante es que un 88% de las cuentas inician con una Proceso Regular y un 12% con el proceso de Demanda anticipada, también llamado proceso “Abogados”.

El problema radica en que de ese 88% de cuentas destinadas al proceso regular, el 60% terminan siendo traspasadas a estudios de abogados por falta de éxito, tomando el global, se tiene entonces que, un 53% de las cuentas fueron designadas “erróneamente”. La poca precisión del modelo actual podría ser explicado por distintos motivos, siendo el principal la antigüedad que ya posee y que la variable objetivo para designar la estrategia para cada cuenta no era la adecuada, es por esto, que se hace necesaria la elaboración de un modelo más robusto, con el fin de optimizar esta asignación y lograr en mejor recupero de saldos de las cuentas en esta etapa.

### 1.2.3. Estimativa valor monetario

En cada asignación mensual sobre que estrategia asignar a cada cuenta se pueden tener dos tipos de errores. El primer error es asignar a estudios de abogados aquellas cuentas que presentarían pagos también mediante agencias de cobranza. En este caso, el cliente logra un acuerdo de pago con el fin de evitar el embargo del vehículo, sin embargo, el proceso legal seguirá avanzando hasta llegar a la etapa de “Notificación”, ya que no sería eficiente en temas operacionales suspender los procesos legales, puesto que, si el cliente no cumple alguna de las cuotas del acuerdo, el proceso legal continuará hasta el embargo. Para los procesos legales involucrados desde la asignación a abogados hasta la notificación se tiene un promedio de costo normalizado de 100 por cuenta.

El segundo error es designar en agencias de cobranza aquellas cuentas que no realizaran un pago mediante esta estrategia. Esto trae como problemas, perdida de la trazabilidad del contrato, un mayor tiempo para alcanzar la etapa de notificación y un mayor devalúo del auto con el paso del tiempo. Se tiene además que a medida que transcurre el tiempo desde el castigo se espera un menor recupero de las cuentas como se puede observar en la figura 4.

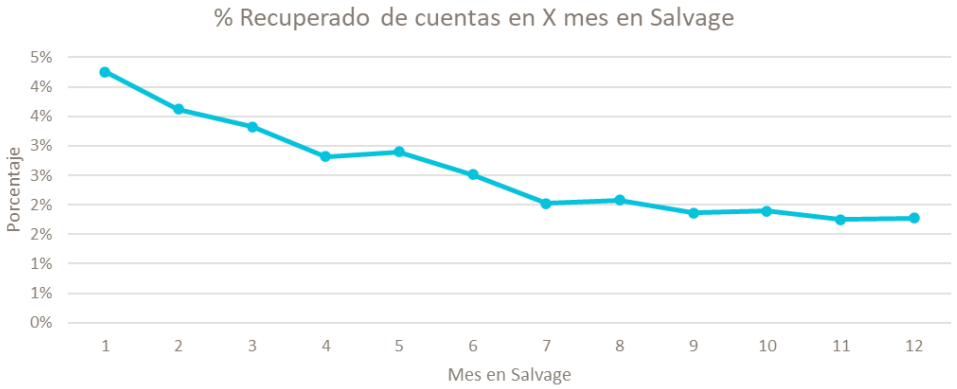



*Figura 4: Porcentaje de recupero promedio por cuenta en “x” mes en Salvage respeto al monto castigado.*

Para estimar el valor monetario del segundo error, se realiza un estudio en cuentas castigadas en junio y julio del 2022. Dentro de estas cuentas castigadas, se establecen una serie de condiciones para seleccionar aquellas de las que se espera un peor comportamiento de pago aplicando el método de cobranza Regular. Una vez definido este grupo, se establece un grupo de control dentro del mismo y un grupo de testeo. Las cuentas en el grupo de control permanecerán en el proceso de cobranza Regular, mientras que las cuentas en el grupo de testeo serán asignadas al proceso de Demanda anticipada. Los resultados de este estudio evidencian una diferencia de recupero del 6% en el grupo de testeo al cabo de 4 meses desde su castigo. Este diferencia porcentual se traduce aproximadamente a un monto normalizado de 285 por cuenta.

# 2. OBJETIVOS

## 2.1. Objetivo general

El objetivo general de este proyecto es el desarrollo de un modelo predictivo de identificación de cuentas riesgosas en cobranza judicial con el fin de mejorar los índices de recuperación de saldos castigados, mediante la determinación de una estrategia especifica de cobranza.

## 2.2. Objetivos específicos

- Identificar las variables que determinen clientes que presentan un mayor porcentaje de pagos una vez han sido demandados.
- Generar modelos de predicción que identifiquen los clientes con mayor propensión a pago mediante abogados.
- Seleccionar el modelo con mejor desempeño en base a métricas de comparación.
- Diseñar estrategias de recupero en base a los resultados del modelo predictivo seleccionado para generar un mayor índice de recupero.

# 3. ALCANCES

El trabajo de título busca predecir la mejor estrategia de cobranza según las características de origen y comportamiento de cada cliente, sin embargo, por temas de tiempo no será posible abarcar, en primer lugar, la implementación del modelo predictivo en reemplazo del modelo actual, y, en segundo lugar, comprobar si existen mejoras en los índices de recuperación al aplicar el nuevo modelo predictivo para seleccionar las estrategias de cobranza especifica por cada cliente.

El alcance del proyecto contempla entonces el entendimiento de la problemática, entendimiento de los datos a utilizar, la preparación, limpieza, creación y selección de datos, creación de modelos predictivos, comparación de modelos predictivos, selección del modelo que presente mejor capacidad predictiva y finalmente un análisis comparativo del modelo seleccionado frente al método de asignación que se está utilizando actualmente (PCA).

# 4. MARCO CONCEPTUAL

## 4.1. Ciencia de datos

La ciencia de datos posee múltiples definiciones, se toma en particular la definida en Oracle, la cual señala que, la ciencia de datos combina múltiples campos, como las estadísticas, la inteligencia artificial, los métodos científicos y el análisis de datos. Esta disciplina abarca desde la preparación de los datos para el análisis, incluida la limpieza, agregación y manipulación de datos para realizar análisis avanzados. Todos estos elementos son utilizados con el fin de descubrir patrones en los datos, y permitir finalmente tomar decisiones respaldadas por estos. Son múltiples los conocimientos y herramientas que son necesarias para utilizar esta disciplina, el proyecto abordara múltiples de ellas, como lo pueden ser, estadística, reconocimiento de patrones, manejo y limpieza de datos, inteligencia artificial y modelos.

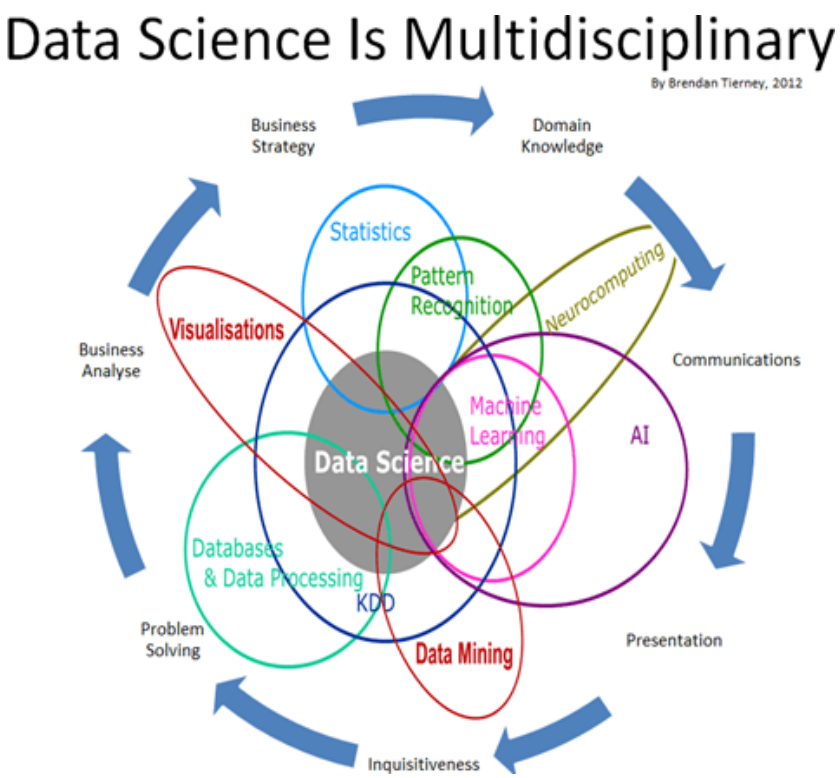



*Figura 5: Data Science is multidisciplinary. Brendan Tierney 2012.*

## 4.2. Aprendizaje supervisado

El aprendizaje supervisado hace referencia a una rama de *Machine Learning*, la cual trabaja con un método de análisis de datos que utiliza algoritmos con una estructura específica, siendo esta, proporcionar al modelo un conjunto de datos y variables junto con una clasificación final y conocida. Los datos en este tipo de algoritmos deben estar

correctamente etiquetados, permitiendo de esta manera un correcto “aprendizaje”. Un ejemplo de problemas que pueden ser abordados con esta técnica son detección de fraude, clasificación de imágenes, predicción del clima, predicciones de demanda, etc. Cabe resaltar la característica en común que tienen estos problemas, la cual es que se conoce el resultado y/o clasificación de cada ocurrencia.

## 4.3. Modelos

Se define como Modelo a un archivo u objeto que se ha entrenado para reconocer patrones en conjuntos de datos. Estos modelos pueden utilizar diferentes algoritmos para identificar estos patrones, como pueden ser:

### 4.3.1. Árbol de decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado, se utiliza para tareas de clasificación y/o regresión. Su estructura consta de un nodo raíz, ramas, nodos internos y nodos hojas como se ve en la figura 6.

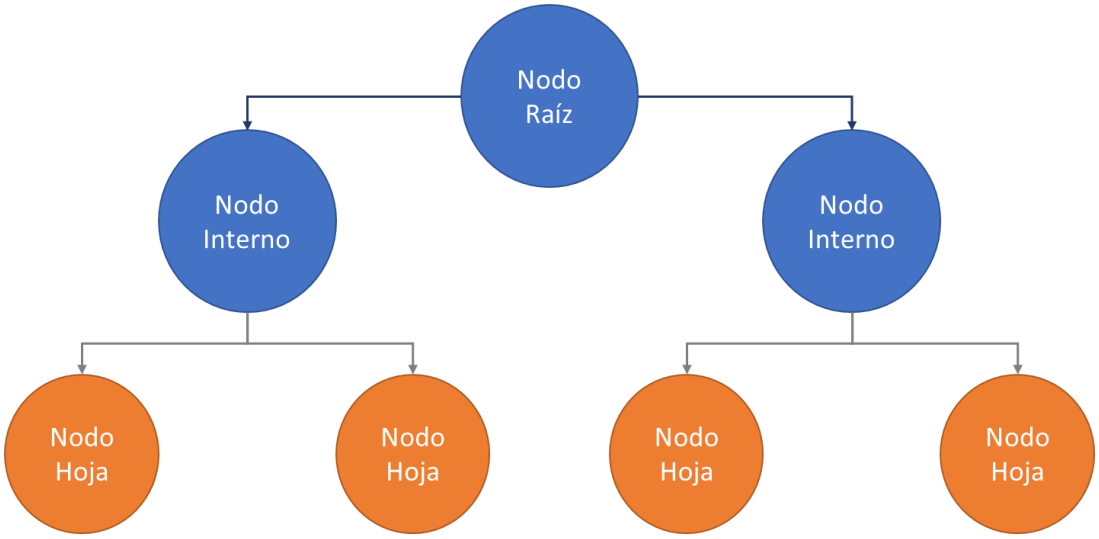



*Figura 6: Árbol de decisión, Elaboración propia.*

Cada uno de los nodos internos realiza evaluaciones sobre el conjunto de datos del modelo, esto se reproduce de manera iterativa hasta llegar a los nodos hoja, que representan todos los resultados posibles dentro de los datos.

### 4.3.2. Random Forest

Random forest es una técnica o algoritmo que consta en la evaluación de decenas, centenas o incluso miles de árboles de decisión, combinando sus resultados para finalmente generar un árbol final más robusto.

### 4.3.3. Generalized Linear Modeling (GLM)

Generalized Linear Modeling es una generalización de los algoritmos de regresión lineal, la cual combina diferentes distribuciones de la función de salida con el fin de explicar la varianza de los valores predichos. En este caso en particular, se trabaja con la función de salida *logit*.

### 4.3.4. Gradient Boosting Machine (GBM)

Gradient boosting machine puede considerarse como una evolución respecto al algoritmo Random Forest, consiste en un ensamblado secuencial de árboles de decisión con el fin de aprender del resultado de árboles previos, buscando corregir el error generado por estos. En este algoritmo el usuario es capaz de seleccionar la extensión de los árboles secuenciales. Este modelo se caracteriza por su rápida velocidad de ejecución (gracias a arboles menos complejos) y una buena precisión de resultados.

### 4.3.5. Redes Neuronales (RN)

Los algoritmos de redes neuronales buscan imitar el funcionamiento de las neuronas del cerebro. Se conforman por nodos, que es una unidad básica que recibe algún tipo de información, y capas, que hace referencia a un conjunto de nodos, múltiples capas conforman la red neuronal, como se puede observar en la figura 7:

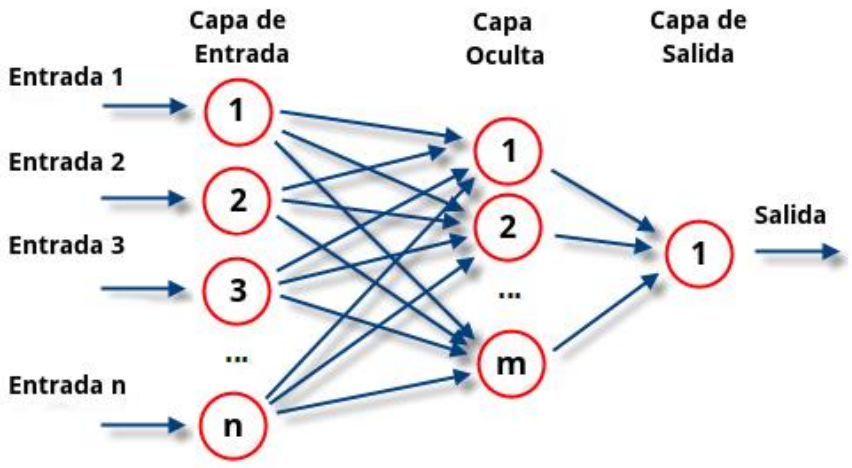



*Figura 7: Que son las Redes Neuronales y sus funciones. Atriainnovation.*

Cada nodo dentro de las capas obtiene un ponderador mediante el aprendizaje del algoritmo, estos modifican el valor de entrada entregándolo como salida al siguiente nodo, finalmente se cuenta con un nodo final de salida que entrega la predicción calculada por la red. Mientras más capas y complejidad poseen las redes, más

operaciones complejas podrá realizar, sin embargo, esto no quiere decir que una red con más capas sea necesariamente mejor que una red de menor dimensión.

## 4.4. Evaluación de los modelos

Dado que se trabaja con más de un modelo, utilizando distintos algoritmos, se hace necesario establecer las métricas de comparación con los que se contrastaran entre sí, con el fin de seleccionar el modelo más adecuado para el problema presentado, en esta sección se detallaran los principales conceptos para tener en cuenta.

### 4.4.1. Accuraccy

El accuracy de un modelo se calcula en base a la matriz de confusión, representando el porcentaje de casos que fueron predichos de manera correcta frente al total de casos pronosticados, mediante la siguiente formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

En donde TP represente los valores positivos predichos correctamente, TN los valores negativos predichos correctamente, FP los valores predichos como positivos que en realidad eran negativos, y FN, que representa valores predichos como negativos que en realidad eran positivos.

### 4.4.2. Recall

Recall hace referencia al total de elementos correctamente predichos como positivos sobre el total de positivos. Se calcula como:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

### 4.4.3. Precision

Hace referencia al número de elementos identificados correctamente como positivos frente al total de elementos predichos como positivos. Se calcula como:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

### 4.4.4. F1 Score

Los componentes principales del Score F1 son la Precision y el Recall. El objetivo es combinar estas dos métricas en una sola. La ventaja que añade este parámetro es que entrega resultados comparables en datos desbalanceados. Se calcula como:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

### 4.4.5. MCC

Matthews correlation coefficient (MCC) es una métrica que toma en consideración los 4 posible casos de la predicción de los datos (TN, TP, FP y FN), este generara un alto valor solo si el modelo es capaz de predecir correctamente tanto la mayoría de los casos positivos como la mayoría de los casos negativos. Se calcula como:

$$MCC = 2 * \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (5)$$

### 4.4.6. Curva ROC

La curva ROC es una curva dentro de un gráfico que muestra en un eje la tasa de falsos positivos y en otra la tasa de falsos negativos, esto permite evidenciar el desempeño general del modelo en distintos puntos de corte de la predicción. El área bajo la curva del grafico (AUC) permite comparar estas curvas entre modelos, mientras mayor sea este valor, el modelo tendrá un mejor desempeño como clasificador.

## 4.5. Balanceo de datos

El balanceo de datos son diferentes métodos que se aplican a un conjunto de datos con el fin de tratar datos desbalanceados, buscando representan de mejor manera las clases minoritarias. Para tratar este desbalanceo existen diferentes técnicas, siendo las principales:

### 4.5.1. Sobremuestreo

Se aplica sobre la clase minoritaria, replicando observaciones de esta para aumentar su representación.

### 4.5.2. Submuestreo

Se aplica sobre la clase minoritaria, consiste en quitar observaciones de esta clase con el fin de equilibrar el total de muestras totales.

# 5. METODOLOGIA Y DESARROLLO METODOLOGICO

La metodología seleccionada para la realización del proyecto corresponde a CRISP-DM.

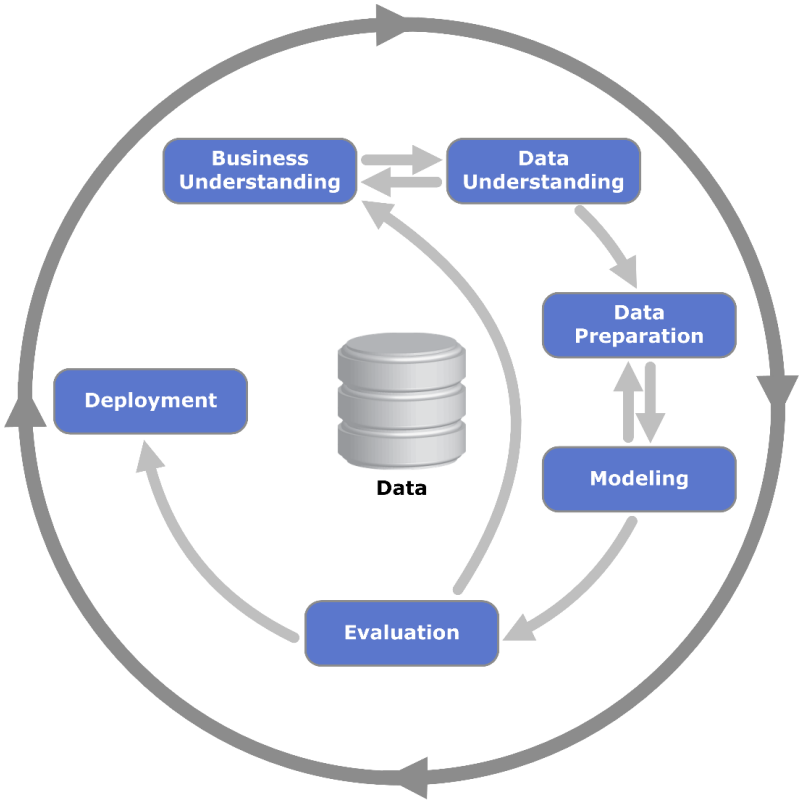



*Figura 8: Metodología CRISP-DM.*

Se eligió este modelo en particular debido a que considera en su primera etapa el entendimiento del negocio. El proyecto se realiza en un sector desconocido, por lo que contar con un modelo que parta de esta base es fundamental. A continuación, se detallará el significado de cada una de las fases de la metodología junto con los resultados y aprendizajes en cada uno de estos.

## 5.1. Entendimiento del negocio

La primera fase de la metodología es el Entendimiento del Negocio. El acercamiento para esta fase fue trabajar directamente en procesos del área y asistiendo a reuniones particulares. Durante estas tareas se aprenden conceptos, flujos y procesos relacionados con el negocio. Se realizan también de manera secundaria, tareas, como elaboración de informes y reportes que involucren trabajar directamente con las bases de datos disponibles dentro de la organización (y que sean posiblemente necesarias para el desarrollo de la investigación). Se realizan también reuniones con el área de Salvage para generar un entendimiento tanto de los dolores del área, como para absorber los

distintos aprendizajes que los trabajadores de esta área ya han adquirido durante su transcurso en la organización. Todo lo mencionado anteriormente va en línea con generar una correcta dirección en la que enfocar el proyecto, con el fin de ser lo más eficiente posible.

Como aprendizaje de lo mencionado anteriormente se tiene que, los contratos en su historial antes de caer a este última etapa denominada Salvage, son capaces de recolectar gran capacidad de información, comenzando en un inicio por la obtención de las variables de originación, siendo estas las variables estáticas con las que se les da inicio a cada contrato, por ejemplo: Características demográficas y personales del cliente, capacidad de pago, modelo del automóvil financiado, porcentaje de pie inicial, monto de las cuotas, meses totales del contrato, entre otras variables las cuales no varían con el tiempo. Posteriormente los contratos a medida que avanzan o retroceden a diferentes Buckets, van recolectando información de comportamiento, como puede ser: Numero de extensiones, días de mora, número de llamadas realizadas en Buckets, puntaje en base a comportamientos anteriores, entre otras.

Cuando las cuentas superan los 120 días de mora, estas son castigadas, por esto se entiende que se vuelve exigible legal y financieramente, el monto total de cuotas restantes hasta el final del contrato desde el momento en que este se castiga. En este momento estas cuentas son traspasadas a Salvage.

El área de Salvage es la encargada de generar recuperos monetarios en todas estas cuentas castigadas. Para lograr esto se cuenta con dos estrategias, siendo estas:

- Proceso regular
- Proceso Demanda anticipada

El proceso regular es menos costoso, pero falla en generar recuperos en los que la segunda estrategia si es capaz de recuperar saldos, sin embargo, la segunda estrategia es más costosa debido a costos relacionados a trámites legales. Es por esto por lo que se hace necesario ser capaz de identificar de manera precisa que estrategia encaja mejor con cada cliente, en base a sus variables de originación y comportamiento. Logrando esto se tendrá un manejo óptimo de recursos, tanto monetarios como humanos.

## 5.2. Entendimiento de la data

La segunda fase corresponde a la familiarización de los datos disponibles, esto se logra mediante la exploración directa de las distintas bases de datos a las que se tiene acceso, con el fin de conocer la cantidad y tipo de variables de cada una de estas, junto también con comprender porque y para que se utiliza cada una. La otra vía de familiarización es, como fue mencionado anteriormente, la realización de tareas que involucre trabajar con estas bases ya sea para visualización de datos, elaboración de presentaciones, generación de métricas clave y participar en el proceso de obtención y publicación de estas.

Para entender los datos con lo que se está trabajando, cabe mencionar que la base consolidada utilizada en el modelamiento proviene de 3 bases diferentes, las cuales son BASE 1, BASE 2 y BASE 3. Se detallan a continuación las bases mencionadas, enunciando las variables que podrían ser relevantes para el proyecto:

### 5.2.1. BASE 1

La BASE 1 , contiene tanto datos de originación del contrato, como datos del comportamiento que tuvo la cuenta antes de caer a Salvage, la granularidad de esta base es mensual, por lo que para generar el número de observaciones útiles para cada cuenta, se toma el primer registro de cada una de estas en el momento en que esta fue castigada, obteniendo así su historial de comportamiento en el momento de ser traspasada a Salvage, omitiendo las observaciones siguientes de esta cuenta, ya que desde el momento que es castigada, se agregan o actualizan variables de comportamiento que podrían ensuciar la predicción final, esto último debido a que la predicción del riesgo o categoría de cada cuenta, se realizara inmediatamente estas cuenta sean castigadas, por lo que tener datos en un periodo posterior a esta fecha estaría sesgando el modelo. Las principales variables de esta base son las siguientes:

| Variable | Tipo | Descripción |
|---|---|---|
| Número de cuenta | Numérica | ID único de cada contrato |
| MES DE CASTIGO | Numérica | Numero de mes de contrato en el cual el cliente fue traspasado a Salvage |
| NUMERO DE PAGOS RESTANTES | Numérica | Numero de cuotas restantes hasta el final del contrato en el momento que el cliente fue traspasado a Salvage |
| MESES DE CONTRATO | Numérica | Numero de meses pactados en el contrato |
| MONTO FINANCIADO | Numérica | Monto total financiado mediante el crédito con GMF |
| PUNTAJE CLIENTE | Numérica | Variable número que establece un puntaje en base a variables de originación |
| CIUDAD | Categórica | Ciudad de origen del cliente |
| REGION | Categórica | Región de origen del cliente |
| CAMPAÑA | Categórica | Campaña bajo la cual el cliente suscribió el contrato |
| EVERX_XMOB | Binaria | Variable binaria que indica si el cliente tuvo X días de morosidad en los primeros X meses de contrato |
| MODELO 1 | Categórica | Modelo del auto financiado |

| Variable | Tipo | Descripción |
|---|---|---|
| MONTO CASTIGADO | Numérica | Monto total castigado del cliente en el momento que fue traspasado a Salvage |
| PERFIL RIESGO IO | Categórica | Categoría de riesgo del cliente en base a variables personales de ingresos y capacidad de pago |
| PIE CATEGORIA | Categórica | Rango de porcentaje de pie del contrato |
| EDAD | Numérica | Edad del cliente |
| NUEVO PUNTAJE COMPORTAMIENTO | Numérica | Puntaje de evaluación de cada cliente. Clientes con un mejor comportamiento poseen un puntaje mayor |
| PIE | Numérica | Porcentaje de pie del contrato |

*Tabla 1: Variables BASE 1*

### 5.2.2. BASE 2

Esta base contiene datos por cada cliente de acciones telefónicas y promesas de pago en los Buckets de morosidad anteriores. La granularidad de esta base es diaria, pero se consolida de forma mensual para reducir la dimensión de la base. Algunas variables relevantes son:

| Variable | Tipo | Descripción |
|---|---|---|
| NUMERO DE CUENTA | Numérica | ID único de cada contrato |
| TOTAL DE LLAMADAS | Numérica | Total de llamadas realizadas por mes a un cliente |
| CONTACTO DIRECTO | Numérica | Total de contactos directos con el cliente. Entendiéndose esto como el total de veces que quien contesto la llamada fue el titular del contrato y no otra persona. |
| PROMESA DE PAGO | Numérica | Total de promesas de pago realizadas por el cliente en las situaciones anteriores. |

*Tabla 2: Variables BASE 2*

### 5.2.3. BASE 3

Esta base contiene todos los pagos realizados por los clientes en etapa de Salvage. La granularidad de esta base es diaria, pero se consolida de forma mensual para reducir la dimensionalidad de la base. Las variables relevantes son:

| Variable | Tipo | Descripción |
|---|---|---|
| NUMERO DE CONTRATO | Numérica | ID único de cada contrato |
| MONTO RECUPERO | Numérica | Monto recuperado en la observación |
| FECHA RECUPERO | Numérica | Fecha de la observación |

*Tabla 3: Variables BASE 3*

### 5.2.4. Descripción de los datos

Con el fin de conocer los datos con los que se está trabajando y generar una caracterización de la base general, se procede a elaborar un análisis descriptivo de una muestra de variables relevantes según el modelo PCA o bien según su relevancia esperada acorde a la naturaleza propia de la variable[3].

**EDAD**

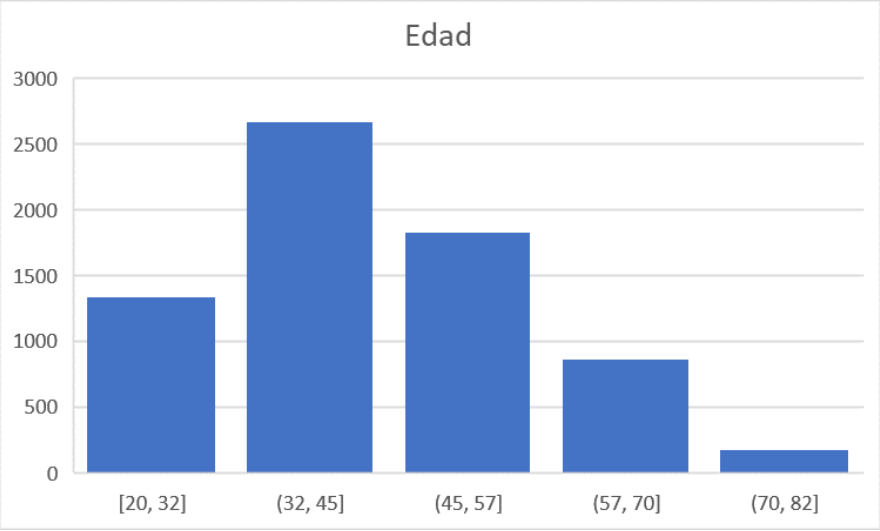



*Figura 9: Edad de los clientes*

Se puede observar que la mayor concentración de datos se encuentra en el rango de 32 a 45 años, aun así, los clientes entre 20 y 32 años sumados a los clientes entre 45 y 57 años presentan una concentración importante.

[3] Con esto nos referimos a que según Juicio experto o experiencia, se espera, por ejemplo, que la variable "Perfil de riesgo del cliente" posea más relevancia relacionada a un comportamiento de pago que la variable "Region".

## PIE CATEGORIA

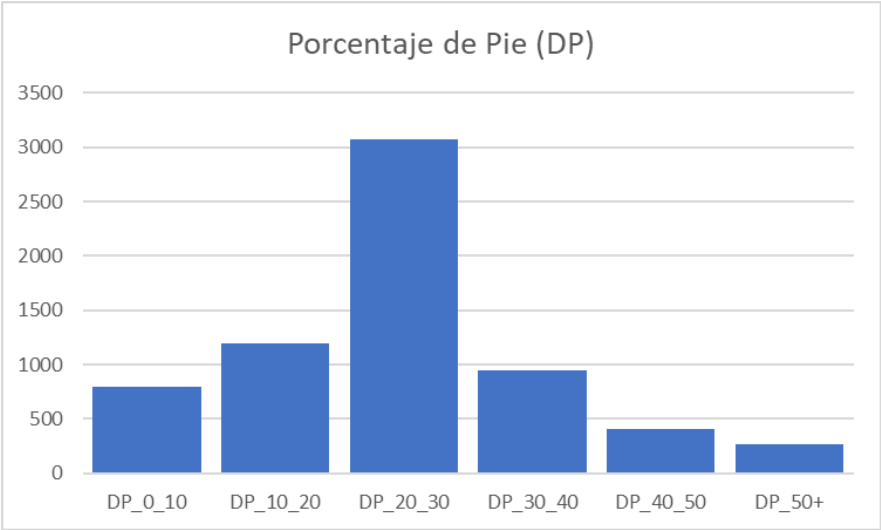



*Figura 10: Porcentaje de pie de los contratos*

El porcentaje de pie se concentra fuertemente en el rango de 20% y 30%. Esto tiene lógica puesto que es el porcentaje de pago "base" de los contratos de GMF, los porcentajes menores son casos menos frecuentes ya que exigen más condiciones y los casos con un porcentaje mayor son de igual manera poco frecuentes, diferenciándose en que la causa de estos es que el mayor porcentaje de pie genera una barrera de entrada, ya que el cliente tendrá que disponer de más dinero inmediato.

## PERFIL DE RIESGO

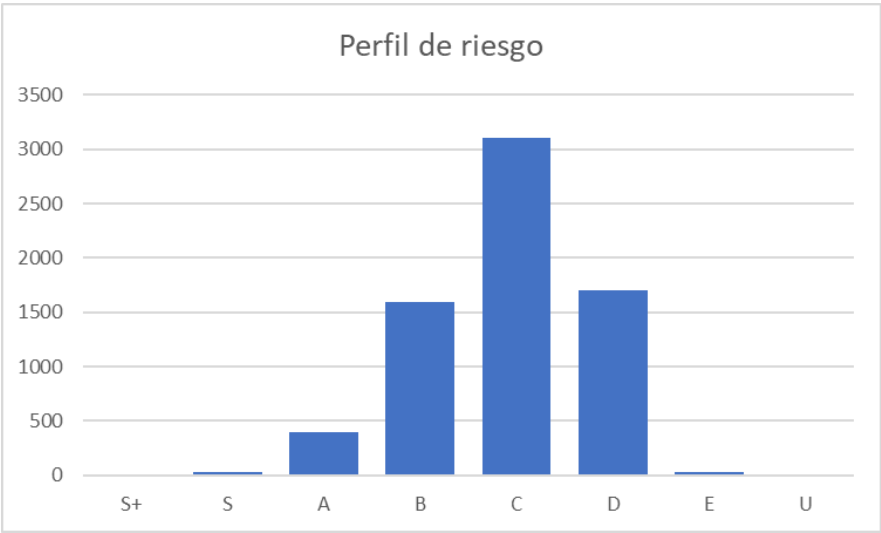



*Figura 11: Perfil de riesgo de los clientes*

El perfil de riesgo de los clientes está concentrado en torno a las categorías B, C y D. Esto se debe a la oferta de vehículos que se pueden financiar mediante los contratos con

GMF, los cuales tienen una alta concentración de vehículos de gama de entrada o media. El orden de las categorías en el grafico se ordena de mejor (S+) a peor (U).

## TOTAL DE LLAMADAS

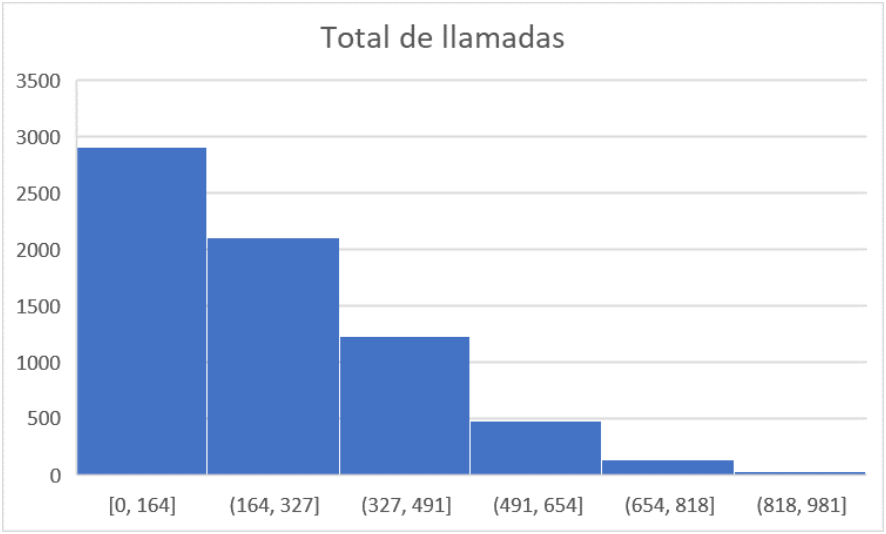



*Figura 12: Total de llamadas de cobranza de los clientes*

El total de llamadas se encuentra concentrado de manera agregada desde los 0 a 327 llamadas (en los últimos 3 meses antes del castigo del contrato). Mientras más alto sea el número, es un indicador de que no se logra contactar al cliente.

## MES DE CASTIGO

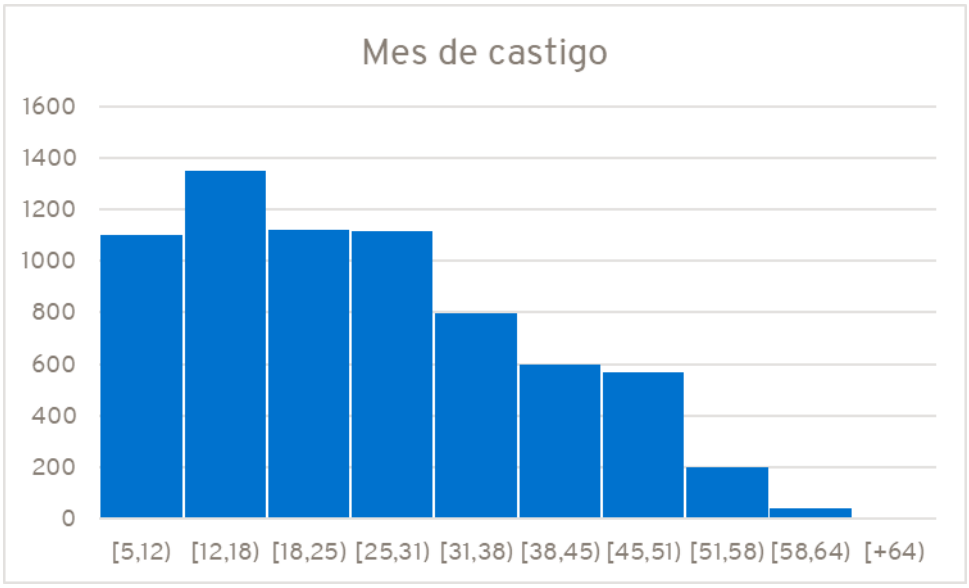



*Figura 13: Mes de castigo de los contratos*

El mes de castigo, se encuentra distribuido de una manera más uniforme. La cifra no es muy indicativa, puesto que hace falta la comparación de este número frente a los meses pactados en el contrato, los cuales varían según el modelo y tipo de contrato que el cliente adquirió.

**MONTO CASTIGADO**

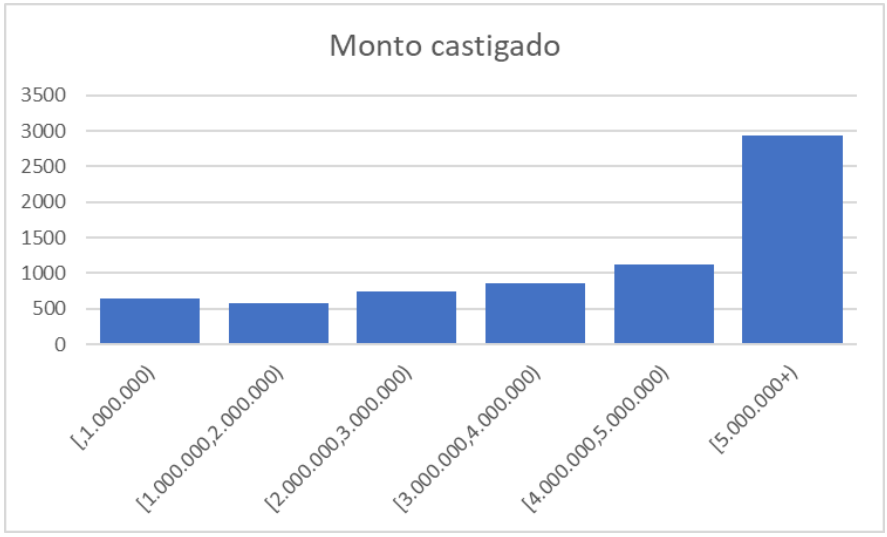



*Figura 14: Monto castigado de los clientes*

El monto castigado se distribuye de manera uniforme en los contratos al ser representados en grupo de rango de un millón de pesos. Los contratos con un monto castigado superior a los 5 millones de pesos presentan una concentración importante.

## 5.3. Preparación de la data

### 5.3.1. Limpieza

Respecto a la limpieza de datos, se eliminan variables que presenten poca o nula variación, también se eliminan columnas con una gran cantidad de NA, ya sea por características de la variable o por falta de completitud de datos al momento de cruzar las tres bases mencionadas anteriormente. También se eliminan registros duplicados de cuentas, lo cual sucede en ocasiones con cuentas que posean dos observaciones en el periodo en el cual fueron castigadas. La última fase relevante de la limpieza de datos corresponde a la aplicación de un filtro, el cual se obtiene de una consulta (en lenguaje SQL) a la base madre de la compañía, en la cual se evidencia en que agencia de cobranza o estudio de abogados se encuentra la cuenta diariamente. El filtro es construido de manera tal que se tiene la fecha y asignación de la primera observación, seguida de todas las fechas y asignaciones en que la cuenta fue traspasada de una agencia a otra, de agencia a abogado, o entre distintas firmas de abogados. Con esto se seleccionan solo las cuentas cuyas primeras asignaciones fueron a agencia. De esta manera se excluyen todas las cuentas que fueron designadas en un principio a abogados.

Esto es relevante ya que el modelo toma en consideración la probabilidad de que una cuenta paga mediante agencias sin necesidad de ser traspasadas a abogados en los primeros tres meses, al incluir cuentas que empezaron su proceso de Salvage en abogados, la muestra podría estar sesgada y predecir de manera errónea. Posterior a la aplicación de este filtro el total de observaciones pasa de aproximadamente 9000 cuentas a un total de 7200 cuentas, con esto se eliminan aproximadamente 1800 cuentas que al ser incluidas en el modelamiento, hubieran sesgado la capacidad predictiva de los modelos.

### 5.3.2. Creación de variables

En base a análisis, reuniones y experiencias pasadas en problemas similares, se identifican posibles variables relevantes que buscan capturar hechos y características clave para cada contrato al momento de estimar la probabilidad de recuperar mediante agencias. En base a los aprendizajes obtenidos de las acciones mencionadas anteriormente se construyen las siguientes variables:

| Variable | Tipo | Descripción |
|---|---|---|
| RATIO CASTIGADO | Numérica | División entre el monto castigado y el monto total financiado del contrato |
| RATIO MADURACION | Numérica | División entre el mes en que el contrato se castiga y los meses pactados del contrato |
| SALDO BAJO | Binaria | Variable binaria que indica si el monto castigado es menor a $1.450.000 . |
| MODELO | Categórica | Modelo abreviado del auto (Sail, Groove, Spark, etc.), se omiten características adicionales con el fin de reducir la cantidad de categorías totales |
| .r_X | Numérica | Monto acumulado recuperado a los X meses en Salvage |
| pc_X | Numérica | Porcentaje recuperado del monto total castigado a los X meses en Salvage |
| RATIO LLAMADAS | Numérica | Porcentaje de contactos directos frente al total de llamadas |
| RATIO PROMESAS | Numérica | Porcentaje de promesas de pago frente al número total de contactos directos |

| Variable | Tipo | Descripción |
|---|---|---|
| PROMESA BIN | Binaria | Variable binaria que indica si el cliente presento o no promesas de pago |
| AVANCE ULTIMOS X | Numérica | Número de veces que el cliente avanzo de Bucket en los últimos X meses |
| RATIO CASTIGO X AVANCE 3 | | Interacción entre las variables RATIO CASTIGO y AVANCE ULTIMOS 3, diferencia clientes con mismo porcentaje de deuda, catalogando con mayor valor a aquellos que no tuvieron intención de pago |
| RETROCESO ULTIMOS X | Numérica | Número de veces que el cliente retrocedió de Bucket en los últimos X meses |
| RC_RMAD | Numérica | Puntaje calculado en base a la multiplicación de las variables RATIO CASTIGO y (1.7 – RATIO MADURACION) |
| Gradiente | Numérica | Pendiente de la recta que une los puntos de PUNTAJE COMPORTAMIENTO 1 mes y 3 meses antes de que el contrato sea castigado |
| PAGOS ULTIMOS X | Numérica | Número de veces que el cliente presento pagos en los últimos X meses (ANTES DE SALVAGE) |

*Tabla 4: Variables creadas*

En esta etapa también se realiza un análisis descriptivo con el fin de conocer las distribuciones y contenido de las variables creadas:

## RATIO CASTIGO

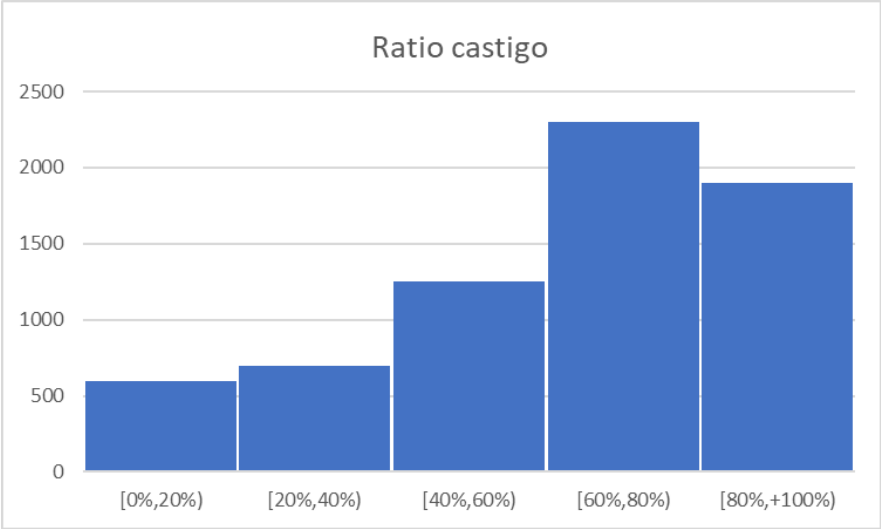



*Figura 15: Ratio Castigo de los clientes*

La variable Ratio castigo presenta una gran concentración de contratos con un ratio castigado mayor al 60%. Los clientes con un gran porcentaje de ratio de castigo han demostrado tener una menor propensión al pago mediante agencias.

## RATIO MADURACION

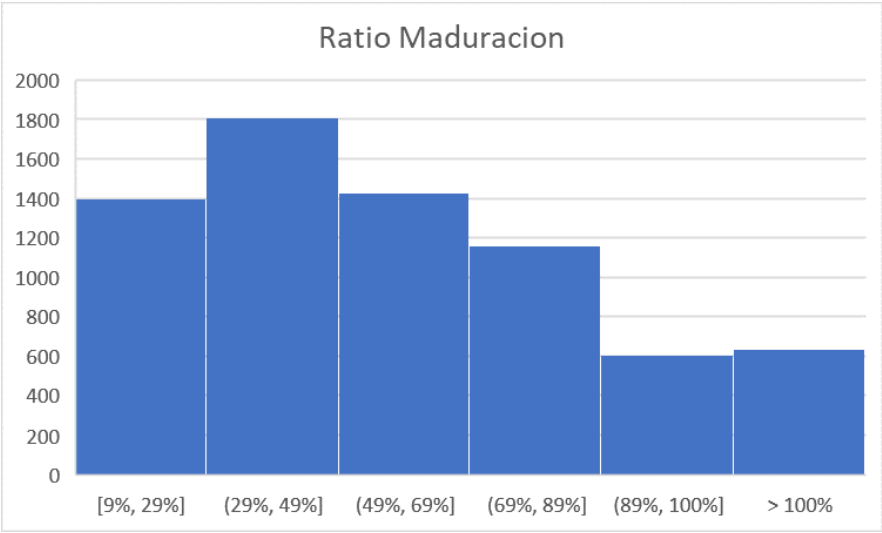



*Figura 16: Ratio de maduración de los contratos*

El ratio de maduración se distribuye de una manera semi uniforme en los rangos normales entre el 0% y el 100%. Los contratos con una maduración mayor al 100% son aquellos en los cuales se logró llegar a un acuerdo de extensión en fechas anteriores al castigo de la cuenta.

## RATIO LLAMADAS

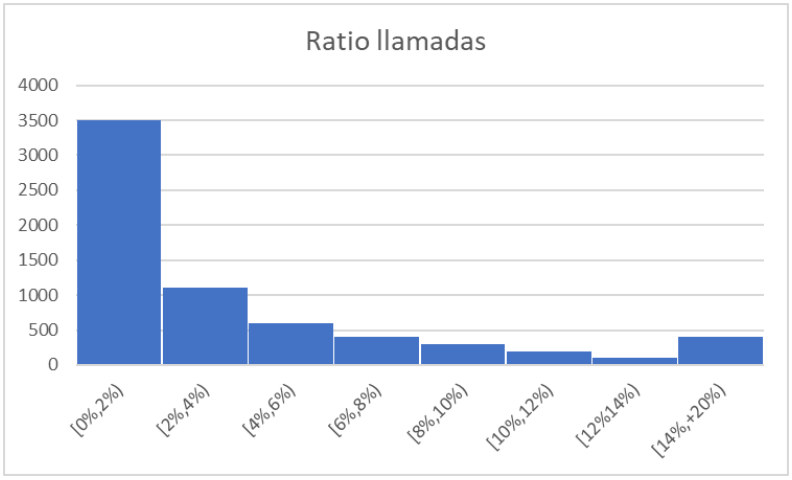



*Figura 17: Ratio llamadas de los clientes*

El Ratio llamadas como se mencionó anteriormente es la razón entre los contactos directos con el cliente (esto siendo, cuando se verifica que quien contesto la llamada es el subscriptor del contrato) frente al total de llamadas. Se puede observar que la gran mayoría de datos se encuentra entre un 0% y 2% de ratio, lo cual es bastante bajo.

## RATIO PROMESAS

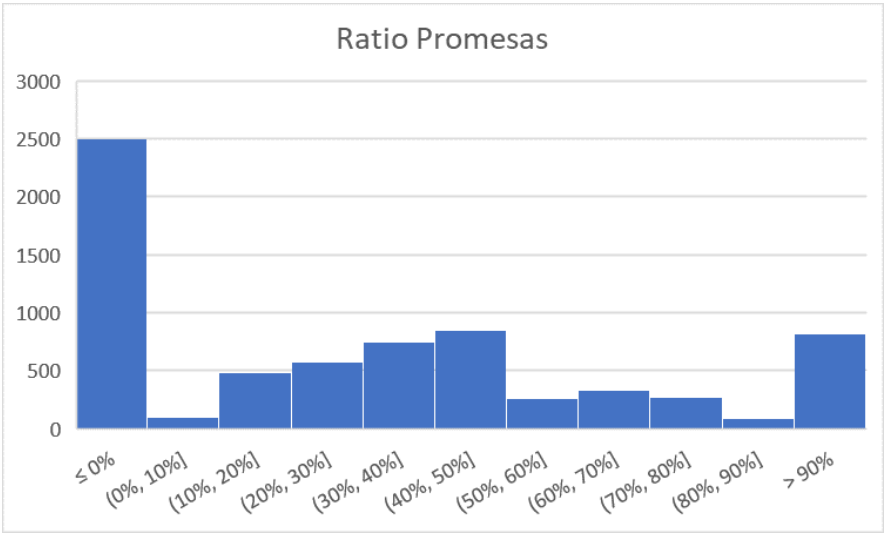



*Figura 18: Ratio promesas de los clientes*

El ratio de promesas como se mencionó anteriormente, corresponde a la razón entre promesas de pago frente al total de contactos directos del cliente. Se puede observar que una gran cantidad de datos se concentra en 0% de ratio de promesas. Aun así, cabe destacar que existe un número no menor de clientes con un ratio mayor al 50%. Estos clientes presentan un mejor comportamiento de pago.

**GRADIENTE**

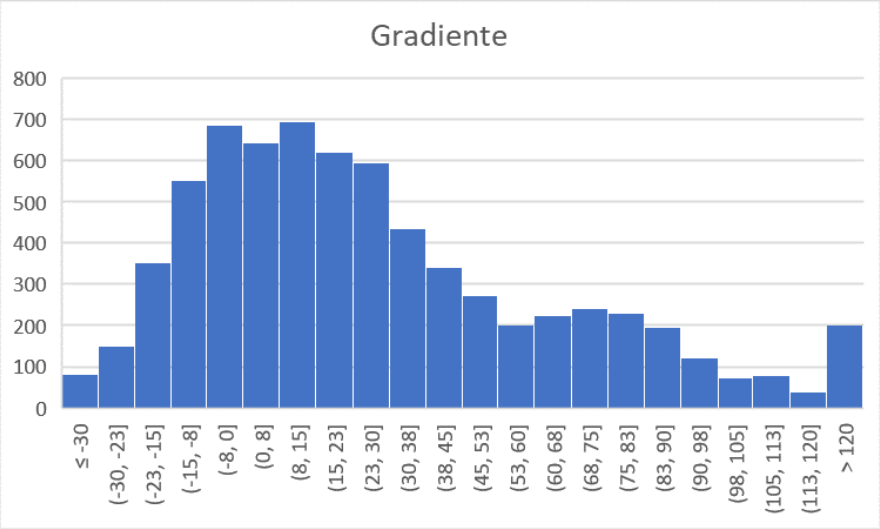



*Figura 19: Gradiente de los clientes*

El gradiente es la pendiente entre el puntaje de comportamiento un mes y tres meses antes de ser castigado el contrato. Si se recuerda que un mayor puntaje de comportamiento representa “mejores clientes”, la pendiente entre estos dos puntos es positiva si el cliente presento un aumento de este puntaje a medida que se acercaba a su fecha de castigo.

## VARIABLE RC_RMAD

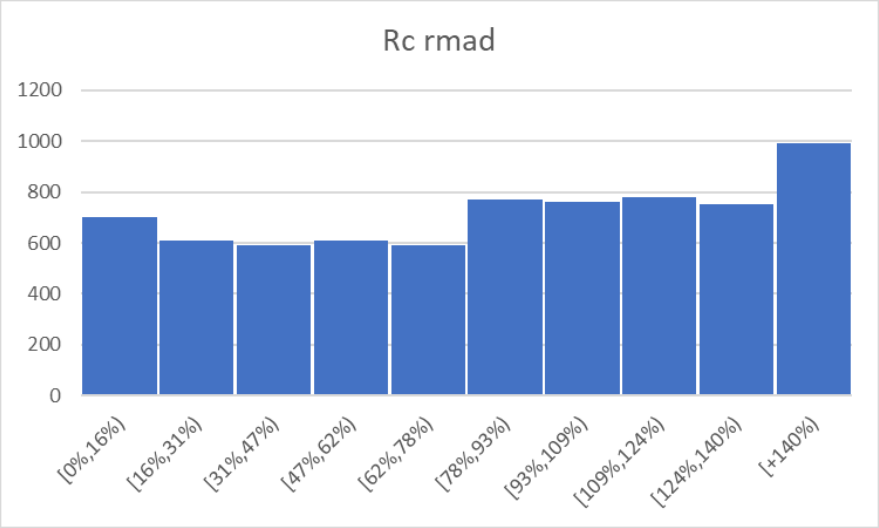



*Figura 20: Variable RC_RMAD de los clientes*

Recordando la variable RC RMAD como la multiplicación entre el ratio castigo y (1.7 – ratio de maduración), mientras más alto sea este porcentaje o ratio, indica clientes con peores características al momento de caer a perdida, ya que estos serán contratos con un gran porcentaje de deuda y un valor bajo de cumplimiento de vida del contrato. La distribución es bastante uniforme y se espera un comportamiento distinto de pago en los extremos de estos valores.

## AVANCE ULTIMOS 3

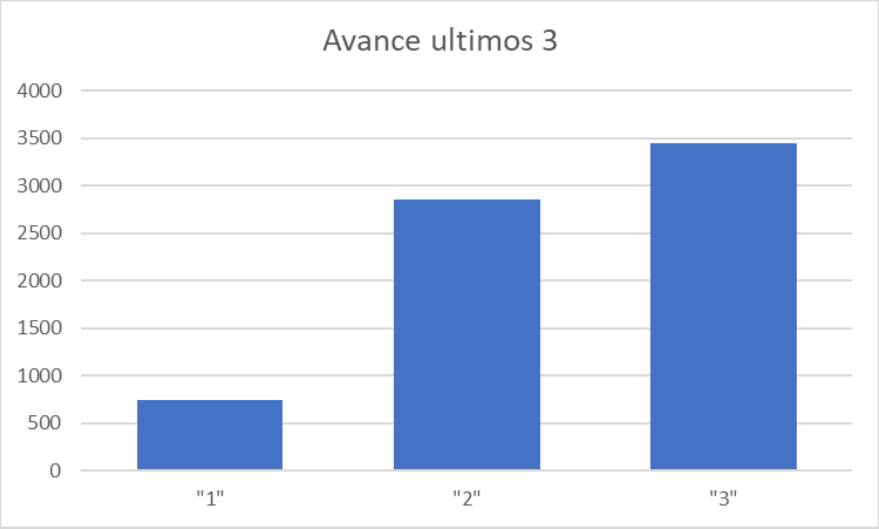



*Figura 21: Avance últimos 3 de los clientes*

Esta variable toma los valores entre 1 y 3, cuando el número es 3, significa que el contrato avanzo 3 Buckets distintos consecutivamente antes de caer a Salvage, si el valor es 1,

significa que solo avanzo una vez, esto se produce cuanto un cliente se encuentra en el Bucket 91-120 y realiza dos pagos en los últimos 3 meses con tal de no caer a esta última etapa.

**RATIO CASTIGO X AVANCE ULTIMOS 3**

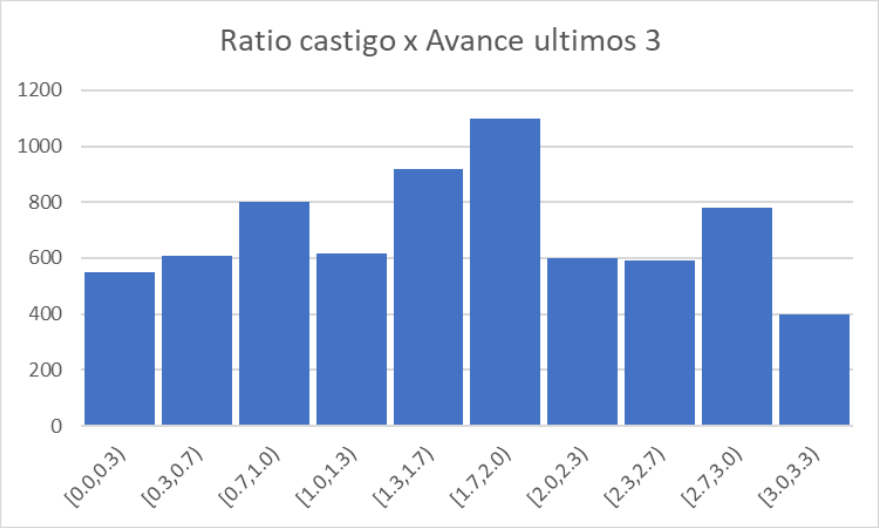



*Figura 22: Ratio castigo x Avance últimos 3 de los clientes*

Esta variable busca diferenciar aquellos clientes con un mismo ratio castigo pero que presentaron mayor intención de pago antes de caer a Salvage, el valor es mayor para aquellos clientes que no realizaron ningún pago en los últimos 3 meses anteriores al castigo. Se espera un comportamiento distinto en los extremos de los rangos.

### 5.3.3. Variable respuesta

La variable por predecir o variable respuesta se construye en base al monto acumulado de las cuentas al tercer mes en Salvage, se define como “No paga” si la cuenta no posee pagos y como “Paga” si posee pagos, esto se debe a que la variable de más interés en predecir aquellas cuentas que no poseerán pagos, ya que es más crítico para el negocio. La distribución de la variable es la siguiente:

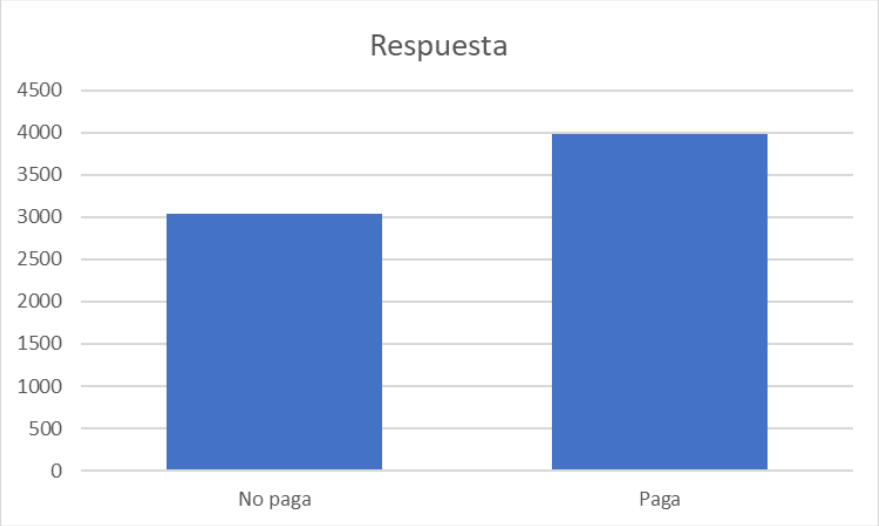



*Figura 23: Distribución de variable Respuesta*

## 5.4. Modelamiento

En esta etapa se seleccionarán variables, se realizará un filtro final a los datos, se dividirán los datos en grupos de entrenamiento, validación y testeo, y finalmente se realizarán 6 modelos para su posterior evaluación.

Los algoritmos por utilizar son:

- Gradient Boosting Machine (GBM)
- Redes neuronales (RN)
- Regresión lineal general (GLM)

Se elaborarán dos modelos con cada uno de estos algoritmos, uno con una base de entrenamiento balanceada y otro con una base de entrenamiento no balanceada.

### 5.4.1. Variables de los modelos

En base a análisis de la etapa anterior se define que las variables a considerar en los modelos son las siguientes:

| Variable | Descripción |
|---|---|
| Respuesta | Variable objetivo que posee un 1 si la cuenta posee pago mediante agencias y un 0 si no |
| CIUDAD | Ciudad de residencia del cliente |
| CAMPAÑA | Campaña bajo la cual el cliente suscribió el contrato |
| MODELO | Modelo del auto financiado |
| TOTAL DE LLAMADAS | Total de llamadas de cobranza realizadas al cliente en los últimos 3 meses antes de la fecha de castigo |
| GRADIENTE | Variable que indica un deterioro o mejora en el puntaje de comportamiento del cliente |
| RATIO CASTIGO | Monto castigado sobre monto financiado del contrato |
| AVANCE ULTIMOS 6 | Número de veces que el cliente avanzo de buckets en los últimos 6 meses |

| Variable | Descripción |
|---|---|
| RATIO LLAMADAS | Numero de contactos directos del cliente sobre el total de llamadas de cobranza realizadas |
| RATIO PROMESAS | Número de promesas de pago sobre el número de contactos directos del cliente |

*Tabla 5: Variables de modelamiento*

### 5.4.2. Consideraciones de los datos

Como se mencionó anteriormente, los datos solo consideran cuentas designadas en agencias, sin embargo, al realizar un análisis del comportamiento de pago de los clientes, se hace necesario aplicar un filtro temporal a los datos. Como se puede observar en la figura 24, el inicio de la pandemia y más fuertemente la promulgación de los sucesivos retiros de fondos de pensiones de las AFP (Administradoras de fondos de pensiones) provocan un significativo ruido en los datos, ya que desde este hito las perdidas bajaron y los recuperos subieron respectivamente a sus mínimos y máximos históricos, es por esto que los datos de cuentas castigadas entre Agosto 2020 y Febrero 2022 no se incluirán en el entrenamiento de los modelos, al no considerar estas observaciones el número de cuentas se reduce a 5937.


*Figura 24: Pérdidas sobre Recuperos*

El grafico representa las *Pérdidas* de cada periodo divido por los *Recuperos* del mismo, normalmente esta proporción se encuentra por sobre 3 en periodos normales, sin embargo, a causa de los hitos nombrados anteriormente, este indicador baja por valores

incluso bajo 1 (periodos en los cuales los montos de recuperos superaron a los montos declarados como pérdidas) en el periodo descrito, este comportamiento es totalmente anormal, por lo que considerarlos dentro del modelo provocaría un ruido erróneo en la relación de comportamiento y pago de deuda de los clientes.

### 5.4.3. Particionamiento de los datos

Los datos comprendidos entre 2018 y Julio 2020 sumados a los datos de enero 2022 se dividirán aleatoriamente en un 80% de base de entrenamiento y un 20% de base de validación de manera iterativa en 10 ocasiones distintas, estas particiones se utilizarán para generar una validación cruzada de los datos con el fin de obtener modelos más robustos. Adicionalmente los datos de marzo y abril del 2022 se utilizan como base de testeo, siendo esta sobre los cuales se medirán y comprarán finalmente los modelos.

### 5.4.4. Balanceo de datos

Para realizar este balance se utiliza una técnica de sobre muestreo de la variable con menos observaciones, la cual es en este caso la variable respuesta 0 (clientes que pagaron mediante agencias de cobranza). El método de sobre muestreo aplicado corresponde a *ROSE* (Random Over Sampling Examples), el cual es un método para balancear muestras de problemas de clasificación binaria.

### 5.4.5. Modelos

Como modelos de predicción se decide trabajar con tres algoritmos, cada uno de estos, como se mencionó anteriormente, se entrenará con una base balanceada y una base no balanceada (base original), los modelos se distribuyen como se evidencia en el siguiente grafico:

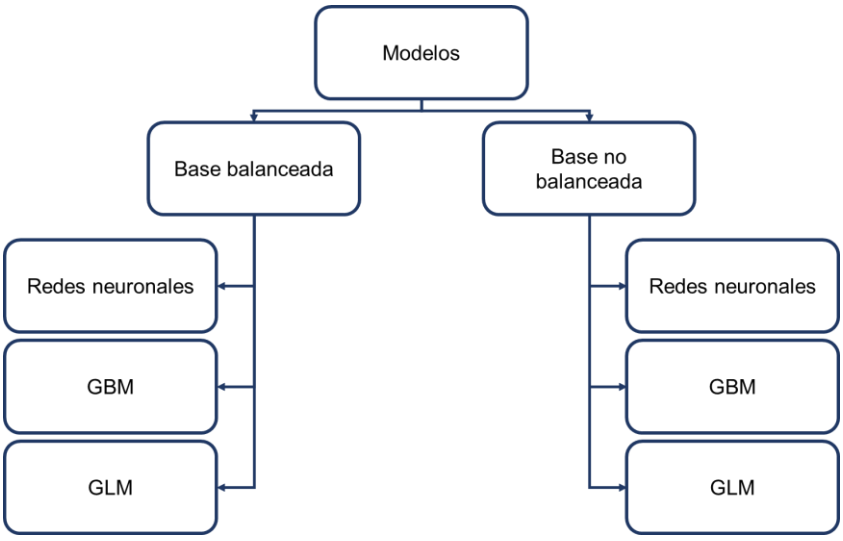



*Figura 25: Modelos a desarrollar y evaluar*

## Modelos con datos balanceados

### *GBM:*

Para este modelo de Gradient boosting machine los mejores parámetros son:

Número de árboles: 50

Profundidad: 5

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 1727 | 1592 |
| | NO PAGA | 383 | 2998 |

*Tabla 6: Matriz de confusión GBM con datos balanceados*

### *RN:*

Para este modelo de RN los mejores parámetros son:

Número de capas: 4

Número de nodos: 170/200/200/2

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 560 | 2759 |
| | NO PAGA | 14 | 3367 |

*Tabla 7: Matriz de confusión Redes neuronales con datos balanceados*

### *GLM:*

Para este modelo de GLM los mejores parámetros son:

Alpha: 0.5

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 2098 | 1221 |
| | NO PAGA | 1144 | 2237 |

*Tabla 8: Matriz de confusión GLM con datos balanceados*

## Modelos con datos no balanceados

### *GBM*

Para este modelo de Gradient boosting machine los mejores parámetros son:

Número de árboles: 50

Profundidad: 5

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 1357 | 1020 |
| | NO PAGA | 356 | 3204 |

*Tabla 9: Matriz de confusión GBM con datos no balanceados*

### *RN*

Para este modelo de RN los mejores parámetros son:

Número de capas: 4

Número de nodos: 170/200/200/2

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 1255 | 1122 |
| | NO PAGA | 421 | 3139 |

*Tabla 10: Matriz de confusión Redes Neuronales con datos no balanceados*

### *GLM*

Para este modelo de GLM los mejores parámetros son:

Alpha: 0.5

Los resultados de la matriz de confusión son los siguientes:

| | | PREDICTED | |
|---|---|---|---|
| | | PAGA | NO PAGA |
| REAL | PAGA | 951 | 1426 |
| | NO PAGA | 620 | 2940 |

*Tabla 11: Matiz de confusión GLM con datos no balanceados*

El desempeño de los modelos en la etapa de entrenamiento y validación cruzada se resume en la siguiente tabla:

| Base | Algoritmo | Accuraccy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|---|
| Balanceada | Gradient Boosting Machine | 0.705 | 0.653 | 0.887 | 0.752 | 0.438 |
| | Redes Neuronales | 0.586 | 0.550 | 0.996 | 0.708 | 0.294 |
| | Genelarized Linear Modeling | 0.647 | 0.647 | 0.662 | 0.654 | 0.294 |
| No balanceada | Gradient Boosting Machine | 0.768 | 0.759 | 0.900 | 0.823 | 0.509 |
| | Redes Neuronales | 0.740 | 0.737 | 0.882 | 0.803 | 0.446 |
| | Genelarized Linear Modeling | 0.655 | 0.673 | 0.826 | 0.742 | 0.251 |

*Tabla 12: Matriz de comparación de modelos en fase de testeo y validación cruzada*

Los resultados en general no presentan una elevada variación de unos a los otros, sin embargo, como se puede observar en las matrices de confusión anteriores, existen casos en que los modelos presentan predicción sobrecargadas a la variable “NO PAGA”, por lo que se hace necesario el análisis del desempeño de estos modelos sobre una base de testeo.

## 5.5. Evaluación y comparación de los modelos

En esta etapa se realiza la evaluación y comparación de los modelos, se evidencia en primer lugar la diferencia de desempeño de todos los modelos realizados y posteriormente se reportan la curva ROC, AUC y la importancia de las variables del modelo que presenta un mejor desempeño en la métrica MCC, no se incluyen representaciones graficas de estas métricas para todos los modelos ya que no se evidencia una diferencia visual relevante (Ver anexos).

### 5.5.1. Métricas de evaluación

Para evaluar y comparar los modelos se calculan las métricas de accuracy, precision, Recall, f1 score y MCC. Los resultados que se muestran a continuación son en base a la base de testeo, la cual corresponde a los meses de marzo y abril 2022 como se mencionó anteriormente, es relevante mencionar la distribución de la variable respuesta en esta

base, la cual es 161 observaciones de “Paga” y 179 observaciones de “No paga”. Los resultados son:

| Modelo | Algoritmo | Accuraccy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|---|
| No Balanceada | Gradient Boosting Machine | 0.597 | 0.719 | 0.385 | 0.502 | 0.242 |
| | Genelarized Linear Modeling | 0.635 | 0.652 | 0.659 | 0.656 | 0.268 |
| | Redes Neuronales | 0.594 | 0.643 | 0.514 | 0.571 | 0.199 |
| Balanceada | Gradient Boosting Machine | 0.585 | 0.583 | 0.749 | 0.655 | 0.163 |
| | Genelarized Linear Modeling | 0.620 | 0.676 | 0.536 | 0.598 | 0.253 |
| | Redes Neuronales | 0.621 | 0.608 | 0.788 | 0.686 | 0.239 |

*Tabla 13: Matriz de comparación de modelos*

Se puede observar que el balanceo de los datos no genera resultados concluyentes a la hora de la predicción de los modelos, ya que en el caso del algoritmo GBM, el que presenta mejor desempeño es el modelo entrenado con los datos no balanceados, por otro lado, en cuanto a las redes neuronales, el modelo entrenado con datos balanceados presenta mejor desempeño a su contraparte no balanceada. Cabe destacar, que, en cada métrica medida, el modelo con mejor desempeño no siempre es el mismo, esto es debido a que tan desbalanceada o balanceada pueda estar la predicción para el punto de corte escogido (seleccionado para el mejor Accuraccy de cada modelo).

### 5.5.2. Curva ROC y AUC

Como medida adicional a las métricas de comparación mencionadas en el punto anterior, se considera también la curva ROC (Receiver Operating Characteristic) y el AUC (Area Under Curve). En esta curva se puede observar la representación gráfica de los ratios de Verdaderos positivos en el eje y los ratios de Falsos positivos en el eje x. En el grafico también se incluye una línea diagonal punteada la cual representa un clasificador aleatorio, el cual por definición posee un 50% de probabilidades de realizar una predicción correcta o incorrecta.

Como se puede observar en la figura 26, la curva ROC presenta un AUC del 65%. Este modelo presenta una mejora frente a un clasificador aleatorio, el cual sería el caso con el actual sistema de asignación de estrategias.

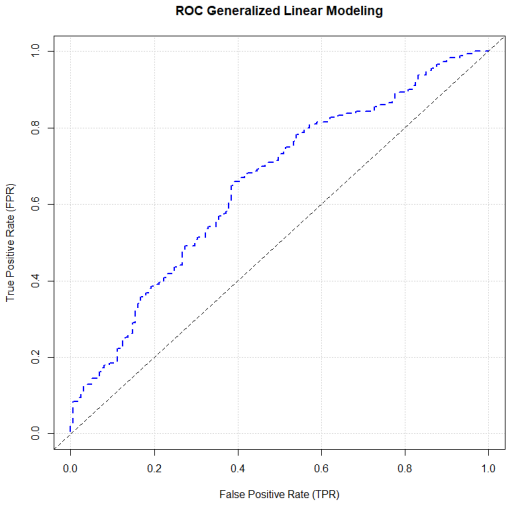



*Figura 26: Curva ROC Generalized Linear Modeling*

### 5.5.3. Importancia de las variables

Posterior a la comparación de los modelos, es necesario conocer cuáles son las variables más relevantes para estos, con el fin de deslumbrar que tipo de variable es más influente a la hora de generar la predicción. Para este análisis solo se trabaja sobre el modelo GLM con datos no balanceados, ya que es el modelo con mejor Accuraccy y MCC. No se muestran las variables relevantes para el resto de los modelos ya que están no difieren de manera importante. La importancia de las variables se muestra a continuación en la figura 27.

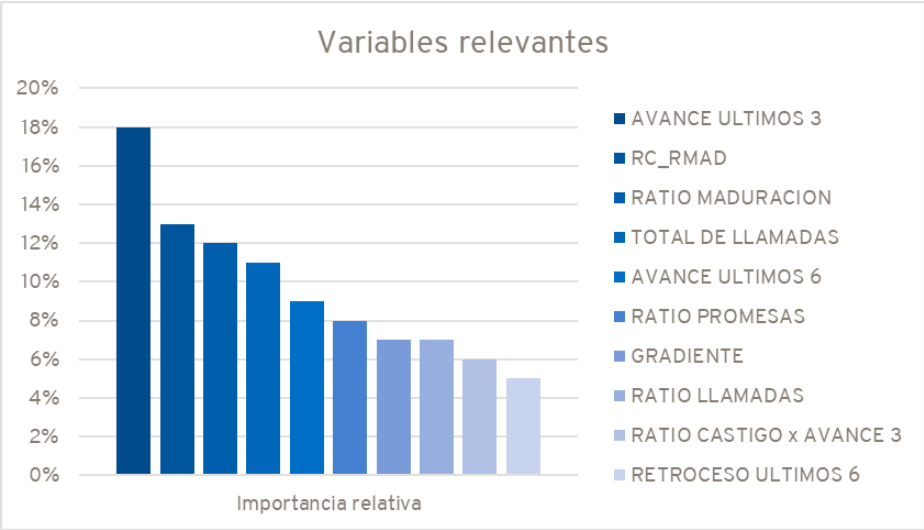



*Figura 27: Variables relevantes*

La variable más relevante es “AVANCE ULTIMOS 3”, recordando esta variable como el número de veces que el cliente avanzo en buckets en los últimos 3 meses. Cabe notar que dejando fuera las variables RC_RMAD y RATIO MADURACION, en su totalidad son variables de comportamiento del cliente, con un gran número que reflejan particularmente 3 meses antes de caer a Salvage. Con esto se puede concluir que este tipo de variables son de gran relevancia a la hora de predecir la variable respuesta del modelo, por sobre las variables de originación.

### 5.5.4. Elección del modelo

El modelo seleccionado como “ganador” deberá ser aquel que presente mejor desempeño en todas las métricas de comparación. Sin embargo, como se puede evidenciar en la Matriz de comparación de los modelos del punto 5.5.1, no hay un modelo que sea dominante frente a los otros de manera absoluta en todos los puntos. Debido al balanceo de los volúmenes de predicción de cada modelo (cuanto porcentaje de casos positivos o negativos predice cada modelo, por ejemplo 80% positivos y 20% negativos o viceversa) hay modelos que presentan el mayor porcentaje de *Precision* pero no presentan el mayor porcentaje en *Accuraccy.* El modelo con mayor F1 Score tampoco es dominante en las otras métricas frente a los modelos competidores. Desde este análisis no es claro de manera inmediata un modelo ganador.

La métrica que se decide como mas relevante para solucionar estas diferencias es la métrica MCC. El puntaje que refleja este indicador es de particular importancia para el problema del punto anterior ya que este toma en consideración los cuatro posibles casos que ocurren al predecir una variable objetivo (TP, TN, FP y FN mencionados en puntos anteriores), dando relevancia a identificar correctamente cada uno de estos casos y penalizando aquellos modelos que presenten sobrecarga en el balance de predicción de la variable respuesta.

El modelo que posee un mayor valor en este criterio es el modelo GLM con datos No Balanceados. Como complemento a poseer el mayor valor MCC el modelo también presenta un mayor desempeño en todos los criterios sobre el *promedio* de todos los modelos evaluados, como se evidencia en la tabla 14.

| Algoritmo | Accuraccy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| Genelarized Linear Modeling | 0,635 | 0,652 | 0,659 | 0,656 | 0,268 |
| Promedio Modelos | 0,609 | 0,647 | 0,605 | 0,611 | 0,227 |

*Tabla 14: Matriz de comparación modelo GLM datos No Balanceados frente a promedio de modelos.*

# 6. CONCLUSIONES

En conclusión, el modelo predictivo seleccionado, es capaz de predecir de mejor manera que el método actual que cuentas presentarán un pago mediante agencias de cobranza en los primeros meses en la etapa de Salvage. Sin embargo, el desempeño del modelo no es el optimo y tiene espacio de mejora. El hecho de que distintos algoritmos entregaran una capacidad de predicción similar incluso al diferenciar estos modelos según incluir o no balanceo de datos en su entrenamiento, habla de que se bordea la mayor capacidad explicativa de la base de datos elaborada para el trabajo de memoria. Este punto es muy relevante ya que gran parte de la calidad de los modelos se basa en cuanta cantidad y calidad de datos hay disponibles para buscar explicar un comportamiento dado.

Para buscar mejorar la capacidad explicativa de los modelos se debe profundizar en la captura de variables que estén relacionadas a predecir si ocurrirá o no un pago de la deuda, todas las variables utilizadas corresponden a variables internas que captura la compañía, dejando fuera aquellas variables externas que si pueden influir en la predicción y que no están siendo consideradas, como pueden ser, datos sobre existencia de deudas de cada persona en entidades distintas a General Motors Financial, numero de vehículos de la persona o variables macroeconómicas como pueden ser la inflación, tasa de desempleo o políticas fiscales y monetarias.

Con respecto a los resultados de los modelos, específicamente en cuanto al tipo de variables mas relevantes para el modelo con mejor desempeño, se concluye que aquellas variables que capturan el comportamiento del cliente desde 3 meses antes de caer a Salvage hasta el momento en que son castigadas, presentan una mayor significancia que las variables “estáticas” o de originación, las cuales pueden ser por ejemplo, Ciudad, Modelo del auto o Compaña bajo la cual el auto fue adquirido. Esto es relevante ya que permite esclarecer una dirección a la cual se debe apuntar a la hora de profundizar la captura de variables, minimizando esfuerzos en la captura de variables de originación y maximizando esfuerzos en la captura de variables de comportamiento.

Respecto a los objetivos específicos del proyecto, el primer objetivo se cumple en la segunda y tercera fase de la metodología CRISP-DM, correspondiente a Entendimiento de los datos y Preparación de los datos, en el cual se realizó un análisis descriptivo de las variables disponibles y se identificaron las variables relevantes a la hora de predecir la realización de un pago en los primeros 3 meses en Salvage.

El segundo objetivo específico, se cumple de igual manera en las fases de Modelado y Evaluación de los modelos. En estas fases se elaboraron 6 modelos, utilizando algoritmos de Redes Neuronales, Gradient Boosting Machine y Generalized Linear Modeling con diferentes datos de entrenamiento.

Finalmente, respecto al tercer objetivo específico, la estrategia de recupero correspondería a asignar estas según la predicción que realice el modelo sobre si cada

cuenta presentara pagos o no, sin embargo, la aplicación de esta estrategia queda fuera de los alcances del trabajo de título.

# 7. TRABAJO FUTURO

Como trabajo futuro se propone la profundización de las variables que capturen el comportamiento del cliente en los momentos anteriores al castigo de la cuenta de este, junto con la incorporación de variables externas que podrían ser relevantes al problema como pueden ser factores de deudas externas o indicadores macroeconómicos relacionados. Para evaluar el impacto de dichos criterios a añadir se deberán contrarrestar los desempeños de los 6 modelos elaborados, enfrentando el grupo entrenado sin estos parámetros contra el grupo entrenado con los parámetros para observar si existe un aumento en la capacidad predictiva.

Con respecto a la evaluación del modelo predictivo frente al modelo PCA, se deberán establecer dos grupos mensualmente en la cuentas que se castiguen cada mes, en el primer grupo, las estrategias de cobranza deberán ser definidas en base al modelo PCA, en el segundo grupo, las estrategias de cobranza deberán ser definidas acorde a la predicción del modelo predictivo con mejor desempeño. Si este modelo efectivamente es superior en cuanto a la asignación efectiva de los métodos de cobranza, el porcentaje de cuentas asignadas a proceso Regular que terminan siendo traspasadas a Abogados por falta de éxito debería ser menor a un 60%, el cual es el porcentaje actual de cuentas que termina siendo traspasadas a Abogados y que iniciaron su proceso en agencias de cobranza según la designación del modelo PCA.  Finalmente, para conocer el impacto monetario, se deberán tomar ambos grupo al cabo de 3 meses y comparar su porcentaje de recupero de deuda total, incluyendo dentro de esta comparación todos los ingresos por pago de deudas y los gastos legales asociados al uso de agencias de cobranza y abogados.

# 8. BIBLIOGRAFIA


[1] Data Science Is Multidisciplinary. (2012, 13 junio). Oralytics. https://oralytics.com/2012/06/13/data- science-is-multidisciplinary/

[2] O. (2021, 15 febrero). Algoritmos de Machine Learning. GraphEverywhere. https://www.grapheverywhere.com/algoritmos-de-machine-learning/#:%7E:text=Los%20algoritmos%20de%20machine%20learning,el%20comportamiento%20de

10. %20los%20datos.

[3] Schröe C., Kruse F., & Marx J. (2020). A Systematic Literature Review on Applying CRISP-DM Process Model. https://doi.org/10.1016/j.procs.2021.01.199

[4] GMF Chile. (s. f.). GM Financial. https://www.gmfinancial.com/es-us/company/careers/locations/latin-america/chile.html

[5] GENERAL MOTORS FINANCIAL CHILE S.A. - Memorias Anuales - CMF. (s. f.). Memoria Anual 2021 GMFChile.

https://www.cmfchile.cl/institucional/mercados/entidad.php?mercado=V&rut=94050000&grupo=&ti
poentidad=RVEMI&row=AAAUvUABfAAAAkgAAX&vig=VI&control=svs&pestania=49

[6] Espinosa-Zúñiga, J. J. (s. f.). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. SCIELO. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000300002

[7] Qué son las redes neuronales y sus funciones. (2019, 22 septiembre). atriainnovation. Recuperado 3 de julio de 2022, de https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

[8] Rodrigo, J. A. (s. f.). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. cienciadedatos. https://www.cienciadedatos.net/documentos/35_principal_component_analysis#:%7E:text=Principal%20Component%20Analysis%20(PCA)%20es,vez%20que%20conserva%20su%20informaci%C3%B3n.

[9] ¿Qué es el aprendizaje supervisado? (s. f.). TIBCO Software. https://www.tibco.com/es/reference-center/what-is-supervised-learning#:%7E:text=El%20aprendizaje%20supervisado%20es%20una,de%20manera%20expl%C3%ADcita%20d%C3%B3nde%20buscar.

[10] Education, I. C. (2021, 30 junio). Supervised Learning. IBM. https://www.ibm.com/cloud/learn/supervised-learning

[11] University of Regina DBD. (s. f.). Confusion Matrix. Uregina. http://www2.cs.uregina.ca/%7Edbd/cs831/notes/confusion_matrix/confusion_matrix.html

[12] Data, S. B. (2019, 27 octubre). Machine Learning: Seleccion Metricas de clasificacion. sitiobigdata.com. https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

[13] Q., Q. (2022, 2 diciembre). ¿Qués es un modelo de aprendizaje automático? Microsoft Learn. https://learn.microsoft.com/es-es/windows/ai/windows-ml/what-is-a-machine-learning-model

[14] Education, I. C. (2021, 26 enero). Random Forest. Microsoft Learn. https://www.ibm.com/cloud/learn/random-forest

[15] Education, I. C. (2021, 26 enero). Random Forest. Microsoft Learn. https://www.ibm.com/cloud/learn/random-forest

# ANEXOS

## ANEXO A: Definiciones

### Recuperos

Hace referencia a los montos de dinero pagados por el cliente en relación con la deuda legal adquirida en la etapa de Salvage.

### Pérdidas

Montos castigados en deuda en un periodo determinado de tiempo. Estas pérdidas suelen ser agrupadas de manera mensual y anual.

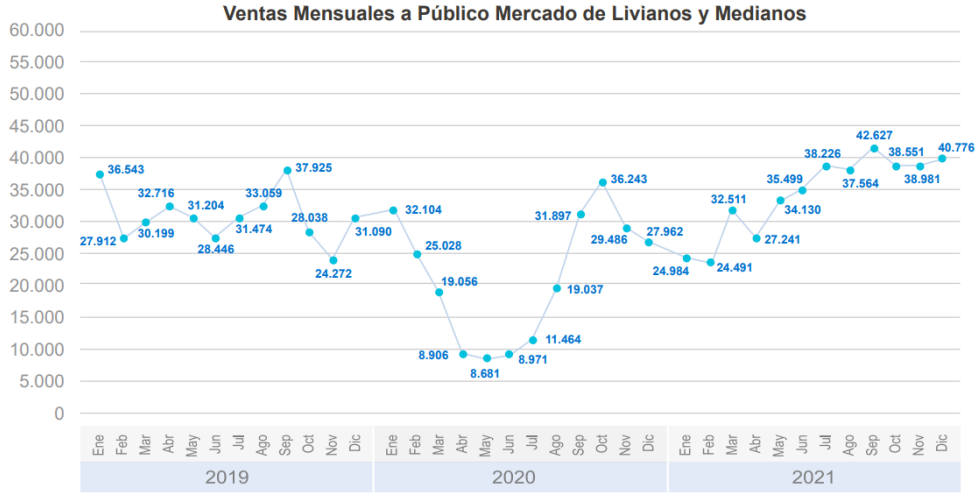
### Buckets

Grupos de cuentas con días de deuda impaga en un intervalo definido. Ej.: Bucket 0 – 30 hace referencia a todas las cuentas que poseen entre 0 y 30 días de mora desde la fecha de vencimiento de la última cuota impaga.

### Castigo de cuenta

El castigo de una cuenta se define como el momento en que esta cumple 120 días de mora, momento en el cual se vuelven exigibles legalmente todas las cuotas restantes hacia el final del contrato por un monto único.

## ANEXO B: Ventas mensuales a público mercado de automóviles livianos y medianos




Fuente: Memoria Anual GMF 2021

## ANEXO C: Directorio General Motors Financial, 31 de diciembre 2021

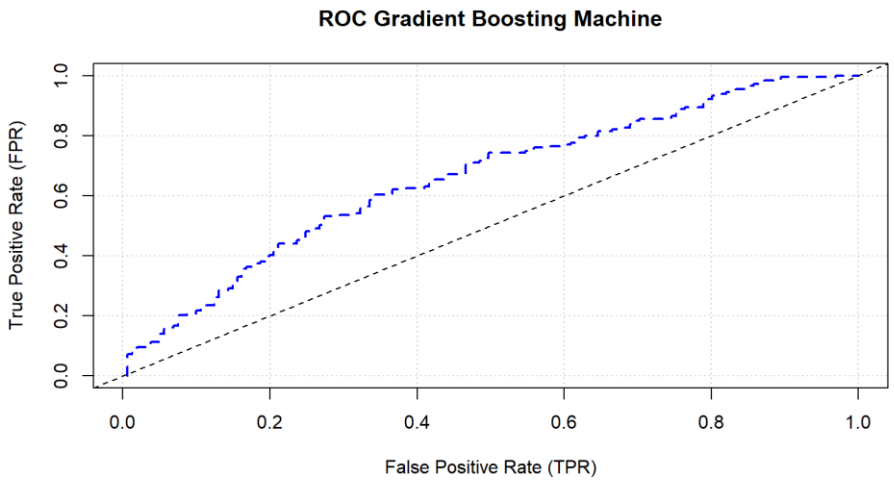
| Nombre | Cargo | RUT | Fecha de Nombramiento | Profesión |
|---|---|---|---|---|
| Paul Holtgreive | Director/Presidente | (*) | 30 de abril de 2021 | Ingeniero Comercial |
| Alessandra Reis Rollo | Director | 26.116.462-0 | 30 de abril de 2021 | Ingeniero Comercial |
| Luis Daniel Moya Alfaro | Director | 23.777.861-8 | 30 de abril de 2021 | Ingeniero Civil |
| Romualdo Andrés Araos Morales | Director | 14.119-618-9 | 30 de abril de 2021 | Abogado |
| Leopoldo Andrés Jeldres Ibáñez | Director Suplente | 13.882.069-6 | 30 de abril de 2021 | Ingeniero Civil |
| Gianina Arias Ormeño | Directora Suplente | 10.437.552-9 | 30 de abril de 2021 | Arqueóloga |
| Tamara Burgos Espinoza | Directora Suplente | 15.096.669-8 | 30 de abril de 2021 | Psicóloga |
| Rubén José Rodríguez Castro | Director Suplente | 23.739.819-k | 30 de abril de 2021 | Ingeniero Comercial |

Fuente: Memoria Anual GMF 2021

## ANEXO D: Curvas ROC

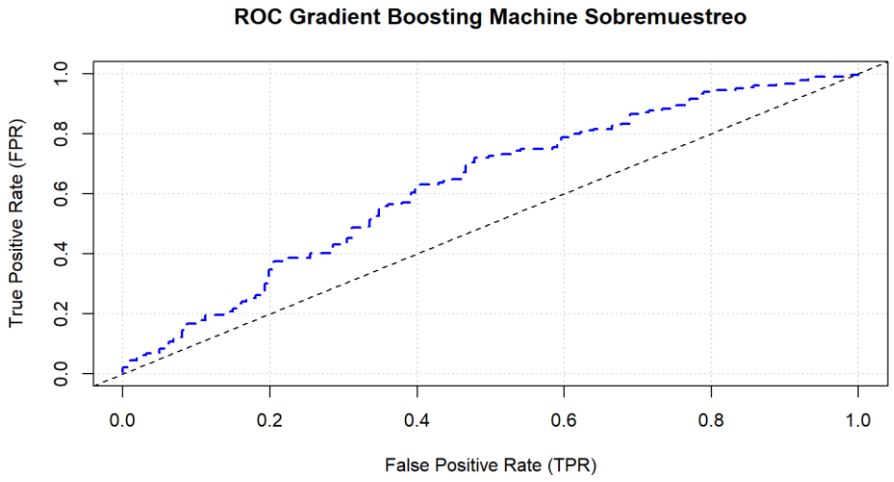
### Modelos Gradient Boosting Machine

#### Gradient Boosting Machine sin balanceo de datos




AUC: 0.65

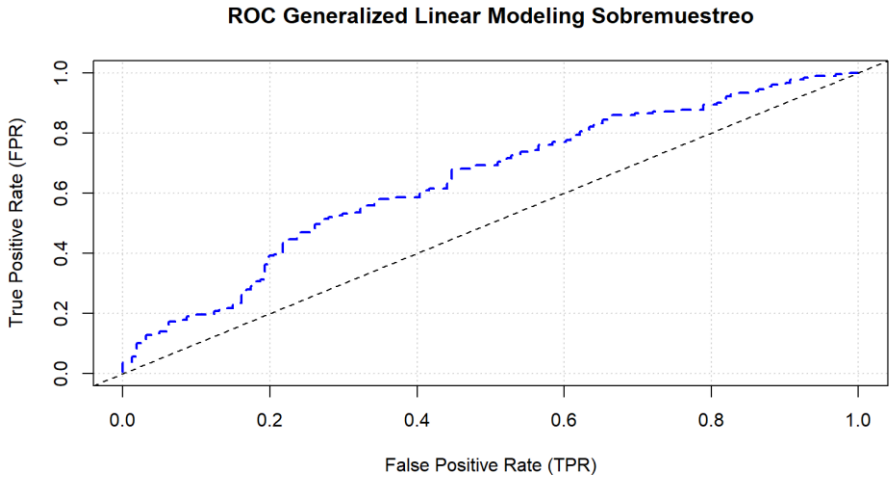
Gradient Boosting Machine con balanceo de datos




AUC: 0.63

## Modelos Generalized Linear Modeling

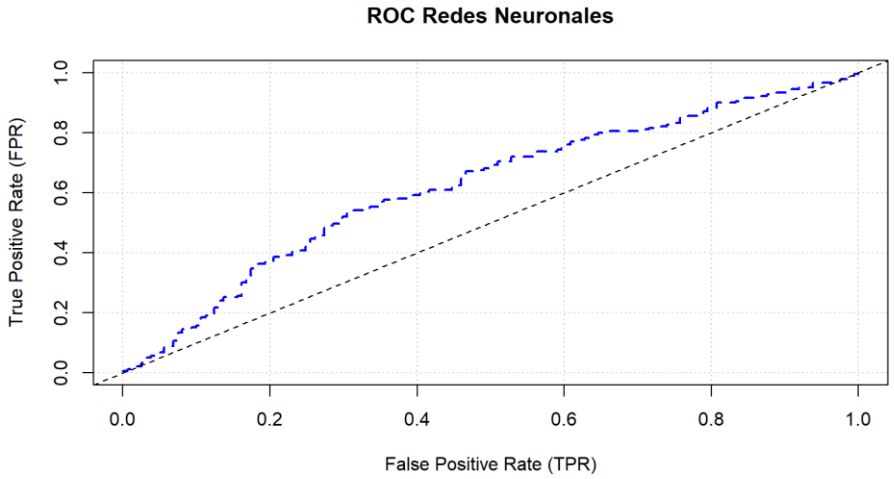
Generalized Linear Modeling con balanceo de datos




AUC: 0.64

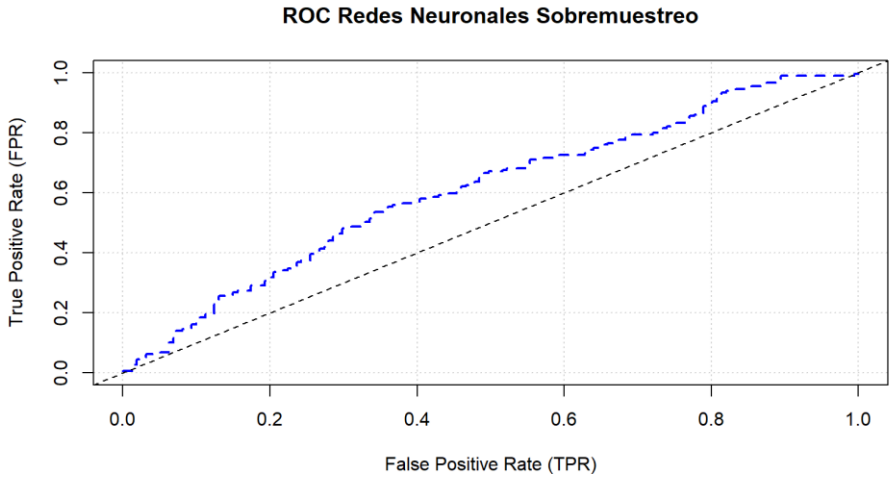
# Modelos Redes Neuronales

## Redes Neuronales sin balanceo de datos



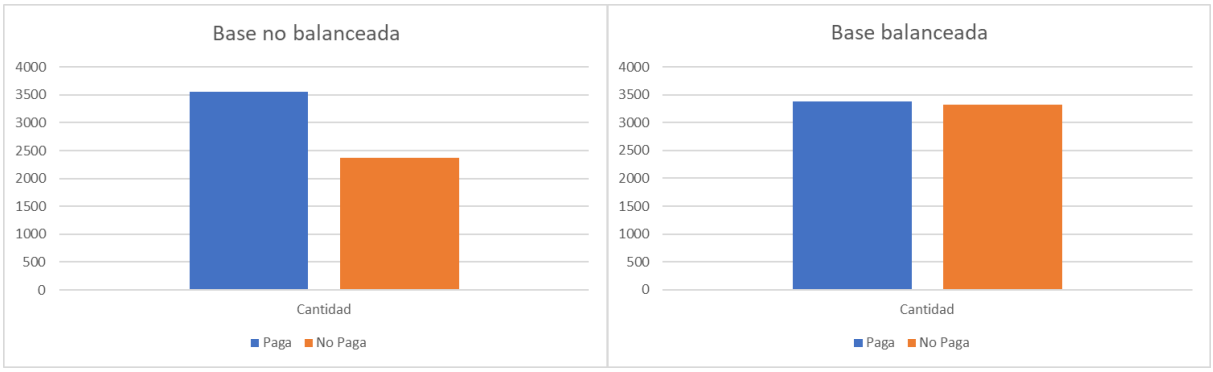

AUC: 0.62

## Redes Neuronales con balanceo de datos




AUC: 0.60

# ANEXO E: Balanceo de datos




Base balanceada vs Base no balanceada