Modelo de Churn para retención de clientes de Seguros Voluntarios

Enrique A. Galvis Moncaleano

Ingeniero Industrial

Trabajo de grado

Faculta de Ingeniería

Maestría en Analítica Aplicada

Director:

Felix Mohr, PhD.

PAGINA DE ACEPTACION

_____

Director Trabajo de Tesis: Felix Mohr, PhD.

_____

Jurado No. 1

_____

Jurado No. 2

_____

Jurado No. 3

Bogotá D. C. mayo 19 de 2023

Agradezco a Dios, por darme la fuerza, la sabiduría y la inteligencia para finalizar con éxito esta etapa de mi vida.

A mi Madre y a mi Hija que, con su apoyo, su amor y su infinita paciencia contribuyeron en gran medida a la culminación de este proyecto.

A mis compañeros que con una gran generosidad me apoyaron y creyeron en mí.

Agradecimiento

Escribo estas pocas líneas para expresar desde lo más profundo de mi corazón el total agradecimiento a todas las personas que con su ayuda colaboraron con la realización del presente trabajo, en especial al profesor Felix Mohr, director de este proyecto, por la orientación, por todo el apoyo recibido, pero sobre todo por la paciencia a lo largo de este tiempo.

A todos, mil gracias.

TABLA DE CONTENIDO

TABLA DE FIGURAS

LISTA DE TABLAS

RESUMEN

La tasa de abandono o Churn se constituye como uno de los más grandes problemas en los negocios masivos de las compañías financieras en Colombia. Toda vez que es mucho más costoso vincular o atraer un nuevo cliente que retener o mantener los ya existentes, se deben crear e implementar estrategias que logren de manera proactiva predecirla y prevenirla, permitiendo a su vez activar campañas comerciales de fidelización y retención de clientes maximizando así la generación de valor.

Con el rápido crecimiento de los sistemas computacionales, las tecnologías de la información asociadas a la transformación digital y la inteligencia artificial, existe una marcada tendencia en las industrias de construcción de sistemas inteligentes y automáticos de gestión para relacionarse con los clientes. Esta tendencia es indiscutible en la actual industria financiera. La predicción de la cancelación de los clientes es una tarea principal de las compañías financieras modernas, conocer el comportamiento futuro de los clientes permite gestionar las relaciones con ellos de manera efectiva y así poder responder a la continua reducción de ingresos en los estados de resultados de las empresas y a la cada vez mayor presión competitiva de los participantes del mercado.

Este trabajo propone desarrollar un modelo para predecir la cancelación de los clientes que adquieren un seguro voluntario y propone el uso de diferentes algoritmos de aprendizaje automático para lograr este fin. Adicionalmente, se utilizan algunas técnicas de minería de datos de uso común para la identificación de clientes que están a punto de abandonar basándose en datos históricos, estos métodos intentan encontrar patrones que puedan identificar posibles abandonos. La explotación de información, el aprendizaje automático y la minería de datos son fundamentales para proporcionar patrones de conocimiento sobre estos clientes.

Palabras Clave: Aprendizaje de máquina, Predicción, Cancelación, Curva ROC

ABSTRACT

The Churn rate is one of the biggest problems in the massive business of financial companies in Colombia. Since it is much more expensive to link or attract a new customer than to retain or maintain existing ones, the strategies must be created and implemented that proactively predict and prevent it, allowing in turn to activate commercial campaigns for customer loyalty and retention, maximizing thus the generation of value.

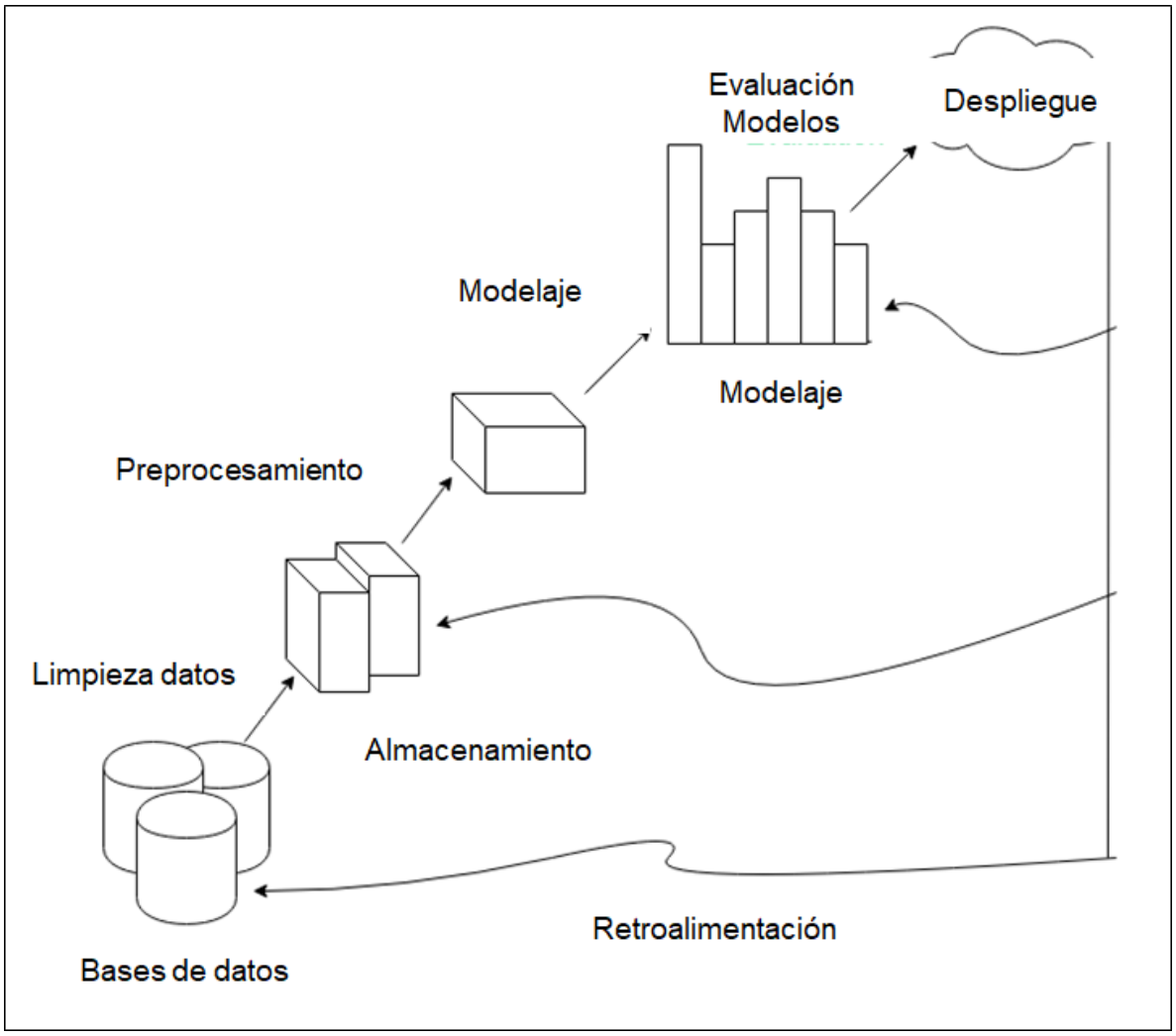
With the rapid growth of computer systems, information technologies associated with digital transformation and artificial intelligence, there is a marked trend in the industries for the construction of intelligent and automatic management systems to interact with customers. This trend is indisputable in today's financial industry. Predicting the cancellation of clients is a main task of modern financial companies, knowing the future behavior of clients makes it possible to manage relationships with them effectively and thus be able to respond to the continuous reduction in income in the income statements of companies and increasing competitive pressure from market participants.

This work proposes to develop a model to predict the churn rate of customers who purchase voluntary insurance and the use of different machine learning algorithms to achieve this end. Additionally, some used data mining techniques are used for the identification of customers who are about to churn based on historical data, these methods try to find patterns that can identify abandonments. The exploitation of information, machine learning and data mining are essential to provide patterns of knowledge about these customers.

Keywords: Machine Learning, Prediction, Churn, ROC curve

RESUMEN GRAFICO

Figura 1 - Esquema de trabajo



INTRODUCCION

El sector financiero se ha convertido en una de las principales industrias en los países desarrollados y por supuesto en Colombia. El progreso técnico y el creciente número de operadores financieros han elevado el nivel de competencia en el País. Estas compañías están trabajando duro para sobrevivir en este mercado competitivo dependiendo de múltiples estrategias para generar mayores ingresos, estas van desde la adquisición de nuevos clientes, el aumento en las ventas para los clientes ya existentes o el aumento del periodo efectivo de duración del cliente con el producto. En Dawkins y Reichheld (1990) se destacaron las ventajas tangibles de retener a los clientes, basado en su experiencia en consultoría y al comparar el resultado de estas estrategias y teniendo en cuenta el valor del retorno de la inversión de cada una de ellas, se puede demostrar que el aumento en la duración del producto o servicio o estrategia de retención es la que presenta un mayor retorno de la inversión, confirmando que retener a un cliente existente cuesta mucho menos que adquirir uno nuevo, adicionalmente puede considerarse mucho más fácil de implementar que la estrategia de conseguir ventas a clientes nuevos.

La retención de clientes es uno de los principales pilares de crecimiento para compañías con un modelo de negocio masivo basado en venta de productos adicionales como lo son los seguros voluntarios. La competencia es dura en el mercado gracias a la cantidad de ofertas, por lo tanto, los clientes son libres de elegir entre muchos proveedores, dentro de esta categoría de productos. Y es aquí donde poder desarrollar un modelo de datos suficientemente robusto que contenga la mayor información posible de aquellos que toman el seguro, sus características sociodemográficas que describan de manera fidedigna a que segmentos de la población pertenecen y la información transaccional como resultado del movimiento y del uso de los productos, genera una ventaja de cara a la obtención de los mejores resultados posibles.

Si la compañía puede predecir de manera efectiva el abandono de los clientes, pueden segmentar e identificar a aquellos que tienen muchas probabilidades de abandonar y así proporcionarles mejores servicios a través de una o varias estrategias comerciales definidas para tal fin. De esta manera, pueden lograr una alta tasa de retención de clientes y limitar la pérdida de ingresos.

Finalmente, en este trabajo se describen los pasos necesarios para la implementación de un modelo predictivo de cancelación, definiendo inicialmente los objetivos que se desean alcanzar, como siguiente paso se detallan los conceptos importantes para comprender la implementación del modelo haciendo énfasis en aquellos que son necesarios para el correcto desarrollo de este problema en específico. Posteriormente se explica la metodología usada basada en el modelo CRISP-DM para minería de datos y finalmente se detallan los resultados obtenidos.

PREGUNTAS DE INVESTIGACION

¿Como implementar un modelo de clasificación binaria que identifique los clientes con probabilidad alta de cancelación?

¿Cuáles son las métricas más adecuadas para evaluar y comparar diferentes modelos de clasificación binaria tales como árboles de decisión, random forest, máquinas de soporte vectorial, naive bayes y XGBoost entre otros, en términos de su capacidad para predecir la cancelación?

¿Cómo se puede interpretar y comunicar de manera efectiva el modelo de clasificación binaria para que los responsables de la toma de decisiones puedan tomar medidas concretas para reducir la cancelación?

MARCO CONCEPTUAL

¿Qué es la tasa de abandono o Churn Rate?

El Churn Rate (o la tasa de abandono de clientes) hace referencia al porcentaje del total de clientes de una compañía que deja de hacer negocios durante un lapso determinado de tiempo. Por lo tanto, es una de las métricas más importantes para conocer el nivel de satisfacción de los clientes con los productos adquiridos o los servicios prestados por la organización.

¿Por qué el Churn Rate es importante?

Se sabe que el costo de adquirir un nuevo cliente es 5 - 6 veces mayor que el costo de retener a un cliente [1]. Como resultado de esto, la compañía puede gastar menos en la adquisición de nuevos clientes, no se necesita invertir tiempo y dinero para convencer a un cliente existente para que elija nuestra compañía sobre los competidores, pues hizo ya esa decisión tiempo atrás. Además, Reichheld and Sasser (1990) en investigaciones posteriores muestran que con solo aumentar (5%) la tasa de retención de clientes se aumenta aproximadamente entre un 25 – 30% el valor presente neto de los clientes en una amplia gama de industrias., como tarjetas de crédito, servicios de automóviles o seguros. Otras investigaciones han también demostrado que el costo de atraer nuevos clientes es mucho más alto que retener a los clientes existentes, Torkzadeh, Chang and Hansen (2006) muestran que este costo puede ser 12 veces mayor.

Causas de cancelación

Las razones por las que un cliente decide irse o cancelar pueden ser personales y únicas, pero generalmente las más comunes son:

Precio: El precio es una de las razones más frecuentes por las que un cliente puede tomar la decisión de abandonar, dado que encuentre un producto o un servicio más económico que el que actualmente posee. Por esto, es muy probable que se den de baja casi de manera inmediata. Por ello, es importante establecer constantemente el valor de la marca y educar al cliente acerca de cómo maximizar los beneficios de los productos o servicios para hacerles sentir que vale la pena continuar con ellos.

Ajuste del producto y mercado: Un ajuste insuficiente entre el producto y el mercado requiere que las ventas y el servicio al cliente se alineen. Si un vendedor se apresura a alcanzar una cuota establecida y no se le incentiva a vender a los clientes que realmente se ajustan al público objetivo, la venta se cancelará en unos cuantos meses.

Experiencia del cliente: Si la atención que recibe un cliente a través de las diferentes áreas de la organización (como los canales de ventas, los equipos de atención a clientes, los Gerentes de cuentas o el soporte técnico) no es positiva, es probable que se vaya. A nadie le gusta sentirse poco atendido, y los clientes desean sentirse bienvenidos y valorados por las personas que prestan un servicio especializado para ayudarles.

Tipos de Cancelación

Existen varias razones por las que un cliente puede tomar la decisión de no continuar con una póliza contratada las más comunes son:

Cancelación por Mora: Esta causal se presenta cuando el cliente no cumple con los compromisos económicos de atender la póliza, por lo tanto, cuando se cumple una mora de 60 días se cancela automáticamente la póliza. Este tipo de cancelación NO se tiene en cuenta en el desarrollo del modelo.

Cancelación por mala venta: Esta situación se presenta cuando el cliente adquiere un seguro producto de una mala práctica de ventas, una mala asesoría o un engaño por parte del asesor comercial. Es una situación grave desde el punto de atención al cliente que puede tener consecuencias disciplinarias para los funcionarios comerciales de la entidad a los que se les atribuye esta práctica. Este tipo de cancelación No se tiene en cuenta en el desarrollo del modelo.

Cancelación por maduración: Cuando se cancela el seguro producto de la terminación del plazo o del prepago del producto financiero asociado. Este tipo de cancelación NO se tiene en cuenta en el desarrollo del modelo.

Cancelación voluntaria: Tomar la decisión de cancelar un seguro es totalmente factible, ya que la situación personal del cliente puede verse modificada, quiere cambiar de compañía o simplemente ha desaparecido el objeto asegurado. Este es el tipo de cancelación que se TIENE en cuenta para el desarrollo del modelo.

Los Algoritmos supervisados

Los algoritmos supervisados de aprendizaje automático pueden aplicar lo aprendido en el pasado a nuevos datos utilizando ejemplos etiquetados para predecir eventos futuros [2]. A partir del análisis de un conjunto de datos de

entrenamiento conocido, el algoritmo de aprendizaje produce una función inferida para hacer predicciones sobre los valores de salida. El sistema puede proporcionar objetivos para cualquier entrada nueva después de una capacitación suficiente. El algoritmo de aprendizaje también puede comparar su salida con la salida correcta e intencionada y encontrar errores para modificar el modelo en consecuencia. Hay dos tipos principales de problemas de aprendizaje automático, llamados clasificación y regresión.

En la clasificación, el objetivo es predecir una etiqueta de clase, que es una elección de una lista predefinida de posibilidades. La clasificación a veces se separa en clasificación binaria, que es el caso especial de distinguir entre exactamente dos clases, y clasificación multiclase que es la clasificación entre más de dos clases. Podemos pensar en clasificación binaria como intentar responder a una pregunta de sí/no. Clasificar correos electrónicos como spam o no spam, detectar si un paciente tiene una enfermedad específica, como diabetes o cáncer, analizar imágenes de plantas y clasificarlas como "enfermas" o "sanas" en función de características visuales, identificar transacciones fraudulentas en sistemas de pagos o tarjetas de crédito o predecir si un cliente va a cancelar uno de nuestros productos o servicios, son algunos de los ejemplos de problemas de clasificación binaria.

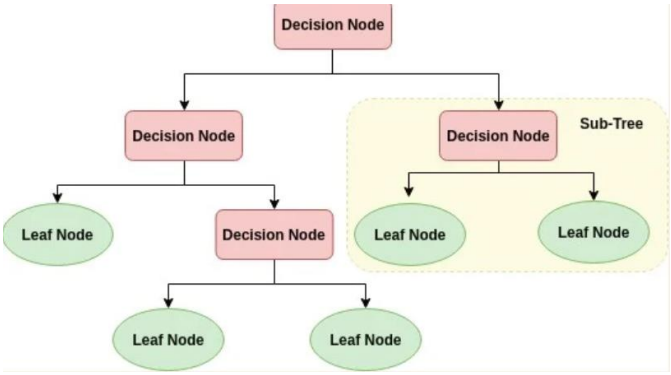
Algoritmos para clasificación Binaria

Existen diferentes algoritmos que se pueden utilizar para construir un modelo de clasificación binaria, y el mejor algoritmo dependerá del conjunto de datos y del problema específico que se esté abordando. Es importante señalar que no hay un algoritmo que funcione mejor en todos los casos, ya que el rendimiento del modelo depende de la calidad de los datos, la cantidad de datos disponibles, la complejidad del problema, entre otros factores. A continuación, se mencionan algunos de los algoritmos de clasificación más populares:

Arboles de Decisión

Los árboles de decisión (ver figura 2), son uno de los algoritmos de Machine Learning más populares, esto se debe a que puede ser fácilmente visible para una persona entender lo que está sucediendo. Un árbol de decisión tiene una estructura similar a un diagrama de flujo donde un nodo interno representa una característica o atributo, la rama representa una regla de decisión y cada nodo u hoja representa el resultado. El nodo superior de un árbol de decisión se conoce como nodo raíz. [3]

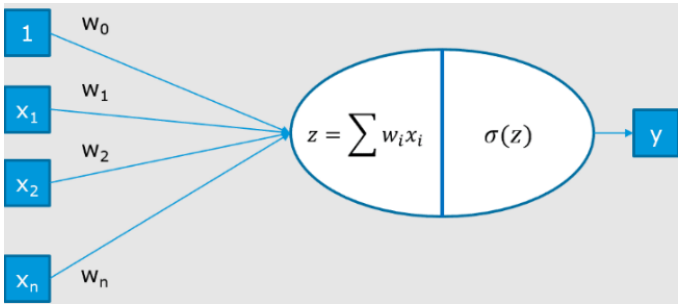
Figura 2 - Visualización de árbol de decisión



Regresión Logística

La Regresión Logística es una técnica de aprendizaje automático para clasificación desarrollada por Cox (1958). Es una red neuronal en miniatura. De hecho, la regresión logística, se trata de una red neuronal con exactamente una neurona. Podemos representar lo que hace una regresión logística en la siguiente figura:

Figura 3 - Representación de regresión logística

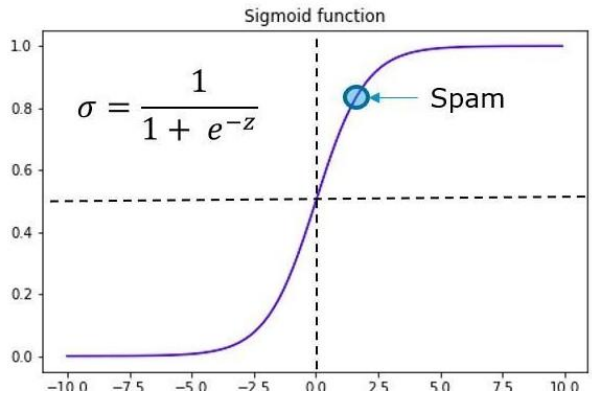


Los valores de **x** corresponden a los distintos atributos de nuestro problema. Por ejemplo, si queremos saber si un correo electrónico es deseado o no deseado los valores de **x** podrían corresponder con cuántas veces aparece cada palabra en un texto. La predicción **y** sería la probabilidad de que el correo sea no deseado. Por lo tanto, la regresión logística tiene dos partes:

- Una combinación lineal (a la izquierda de la neurona).
- Aplicación de la función logística (a la derecha de la neurona)

Así que todas las entradas se combinan con una línea con los coeficientes w. Y luego se aplica la función logística (también llamada sigmoide) al resultado.

Figura 4 - Función logística



Características de la función logística

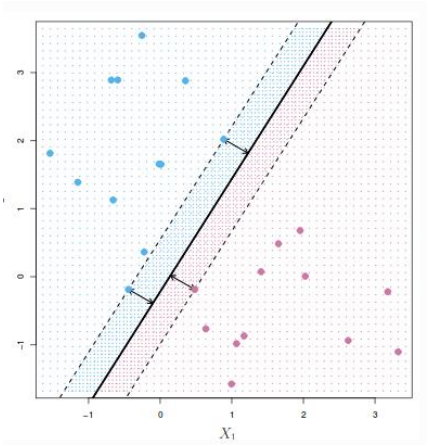
- Está acotada entre 0 y 1. Su valor mínimo es 0 y el máximo es 1

- Podemos interpretar sus resultados como probabilidades (por ejemplo, una probabilidad 0.85 de que el correo sea spam).
- Para problemas de clasificación binaria, podemos suponer que los valores menores de 0.5 corresponden a la clase 0 y los superiores a 0.5 a la clase 1.

Máquinas de soporte vectorial (SVM)

El método de clasificación-regresión " Support Vector Machines – SVM" fue desarrollado por (Cortes and Vapnik, 1995), dentro de campo de la ciencia computacional. Si bien originariamente se desarrolló como un método de clasificación binaria, su aplicación se ha extendido a problemas de clasificación múltiple y regresión. SVM ha resultado ser uno de los mejores clasificadores para un amplio abanico de situaciones, por lo que se considera uno de los referentes dentro del ámbito de aprendizaje estadístico y machine learning.

Figura 5 - SVM Hiperplano



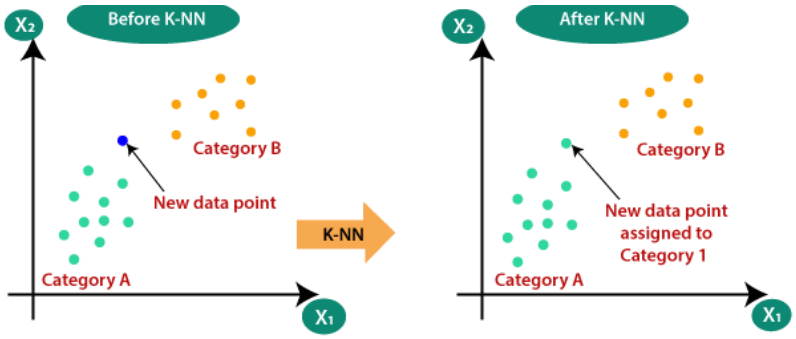
La solución a este problema consiste en seleccionar como clasificador óptimo al que se conoce como hiperplano óptimo de separación, que se corresponde con el hiperplano que se encuentra más alejado de todas las observaciones de entrenamiento. Para obtenerlo, se tiene que calcular la distancia perpendicular de cada observación a un determinado hiperplano. La menor de estas distancias

(conocida como margen) determina como de alejado está el hiperplano de las observaciones de entrenamiento. El maximal margin hyperplane se define como el hiperplano que consigue un mayor margen, es decir, que la distancia mínima entre el hiperplano y las observaciones es lo más grande posible. Aunque esta idea suena razonable, no es posible aplicarla, ya que habría infinitos hiperplanos contra los que medir las distancias.

K-Vecinos más cercanos (K-NN)

K-Nearest-Neighbor es un algoritmo basado en instancia de tipo supervisado de Machine Learning que puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Al ser un método sencillo sirve esencialmente para clasificar valores buscando los puntos de datos “más similares” (por cercanía) aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basado en esa clasificación. Es un método que simplemente busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de los datos que le rodean. [4]

Figura 6 - Visualización de un modelo de KNN

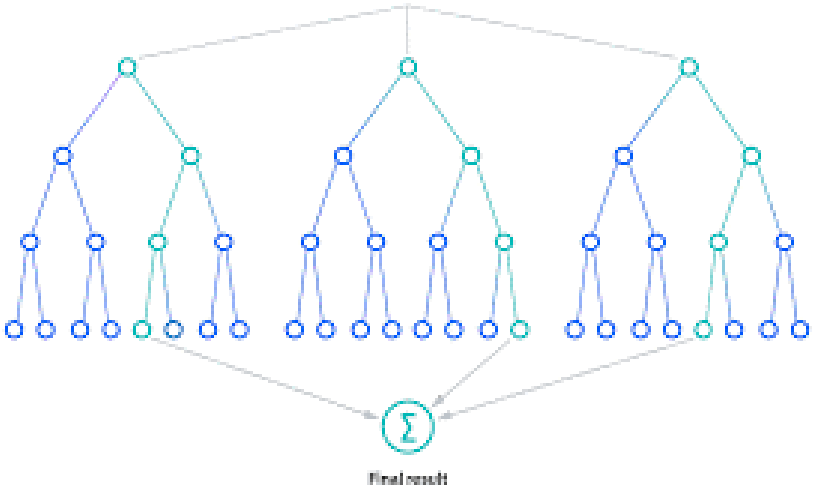


Random Forest

El Random Forest es un algoritmo de machine learning de uso común registrado por Leo Breiman y Adele Cutler, que combina la salida de múltiples árboles de decisión para alcanzar un solo resultado. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión. Dependiendo del tipo de problema, la determinación de la predicción variará. Para una tarea de regresión, se promediarán los árboles de decisión individuales, y para una tarea de clasificación, un voto mayoritario, es decir, la variable categórica más frecuente, arrojará la clase predicha.

Dado que el algoritmo Random Forest puede manejar tareas de regresión y clasificación con un alto grado de precisión, es un método popular entre los científicos de datos. El agrupamiento de características también convierte al clasificador de Random Forest en una herramienta eficaz para estimar los valores perdidos, ya que mantiene la precisión cuando falta una parte de los datos.

Figura 7 - Visualización modelo Random Forest)



XGBoost

XGBoost es un método de aprendizaje automático supervisado para clasificación y regresión, XGBoost es la abreviatura de las palabras inglesas "extreme gradient boosting" (refuerzo de gradientes extremo). Este método se basa en árboles de decisión y supone una mejora sobre otros métodos, como el bosque aleatorio y refuerzo de gradientes.

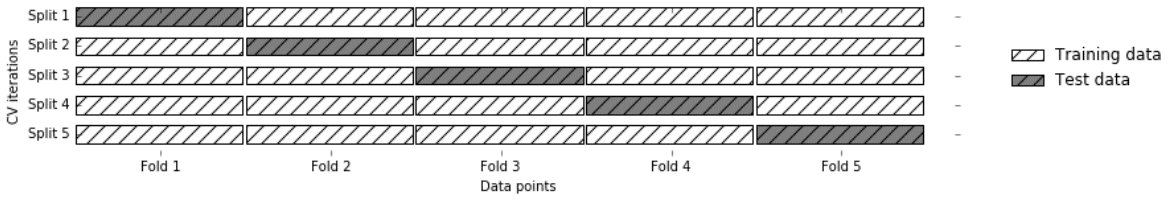
Los modelos de aprendizaje automático pueden ajustarse a los datos individualmente o combinarse con otros modelos, creando un conjunto. Un ensemble es una combinación de modelos individuales simples que juntos crean uno más potente. El boosting de aprendizaje automático es un método que crea este tipo de conjuntos, comienza ajustando un modelo inicial (en nuestro caso un árbol de regresión o clasificación) a los datos. A continuación, se construye un segundo modelo que se centra en predecir con exactitud las observaciones que el primer modelo predijo mal. Se espera que la combinación de estos dos modelos sea mejor que cada uno de ellos. Este proceso de refuerzo se repite varias veces, y cada modelo sucesivo intenta corregir las deficiencias del conjunto refuerzo combinado que contiene todos los modelos anteriores.

El refuerzo de gradiente es un tipo de refuerzo de aprendizaje automático. Se basa en la intuición de que el mejor modelo siguiente posible, cuando se combina con los modelos anteriores, minimiza el error de predicción global. La idea clave es establecer el peso de cada observación para este próximo modelo con el fin de minimizar el error. El nombre de boosting de gradiente surge del hecho de que cada peso se establece en función del gradiente del error con respecto a la predicción. Cada nuevo modelo da un paso en la dirección que minimiza el error de predicción, en el espacio de predicciones posibles para cada observación.

Validación cruzada

La validación cruzada es un método estadístico para evaluar el rendimiento de un modelo que es más estable y completo que usar la división en un conjunto de entrenamiento y otro de prueba. En la validación cruzada, los datos se dividen repetidamente y se entrenan varios modelos. La versión más utilizada de validación cruzada es la validación cruzada de k-folds, donde k es un número especificado por el usuario, generalmente 5 o 10. Al realizar una validación cruzada de cinco veces, los datos primero se dividen en cinco partes aproximadamente de igual tamaño, llamados pliegues. A continuación, se entrena una secuencia de modelos. El primer modelo se entrena usando el primer pliegue como conjunto de prueba, y los pliegues restantes (2–5) se usan como conjunto de entrenamiento. El modelo se construye utilizando los datos de los pliegues 2 a 5 y luego se evalúa la precisión en el pliegue 1. Luego se construye otro modelo, esta vez usando el pliegue 2 como conjunto de prueba y los datos en los pliegues 1,3,4 y 5 como conjunto de entrenamiento. Este proceso se repite utilizando los pliegues 3, 4 y 5 como conjuntos de prueba. Para cada una de estas cinco divisiones de datos en conjuntos de entrenamiento y prueba, calculamos una métrica de desempeño. El proceso se ilustra en la figura 8:

Figura 8 - Visualización validación cruzada (K-Folds)



Métricas de evaluación

En el campo de la inteligencia artificial y el aprendizaje automático una matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, o sea en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje con los datos.

En la vida real, hay dos posibles verdades: lo que se está probando es verdadero o no. (Por ejemplo, la persona está enferma o no lo está). Debido a esto, también hay dos resultados de prueba posibles: un resultado de prueba positivo (la prueba predice que la persona está enferma o no). Estas cuatro opciones se resumen en la figura 8, debido a que surgen dos posibles valores reales y dos posibles valores de predicción o predictivos.

Figura 9 - Matriz de Confusión

Verdadero positivo (VP): El valor real es positivo y el modelo predijo también que era positivo. O bien una persona está enferma y el modelo así lo demuestra.

Verdadero negativo (VN): El valor real es negativo y el modelo predijo también que el resultado era negativo. O bien la persona no está enferma y el modelo así lo demuestra.

Falso negativo (FN): El valor real es positivo, y el modelo predijo que el resultado es negativo. La persona está enferma, pero el modelo dice de manera incorrecta que no lo está. Esto es lo que en estadística se conoce como error tipo II.

Falso positivo (FP): El valor real es negativo, y el modelo predijo que el resultado es positivo. La persona no está enferma, pero el modelo nos dice de manera incorrecta que si lo está. Esto es lo que en estadística se conoce como error tipo I. A partir de estas 4 opciones surgen las métricas de la matriz de confusión, así:

La Exactitud ("Accuracy"): se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Se representa como la proporción de resultados verdaderos (tanto verdaderos positivos (VP) como verdaderos negativos (VN)) dividido entre el número total de casos examinados (verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos).

$$\frac{(VP + VN)}{(VP + FP + FN + VN)}$$

La Precisión ("Precision"), se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos).

$$\frac{VP}{(VP + FP)}$$

La Sensibilidad (“Recall” o “Sensitivity”), también se conoce como Tasa de Verdaderos Positivos (True Positive Rate) o TP. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo. Se calcula así:

$$\frac{VP}{(VP + FN)}$$

En el área de la salud se dice que el “Recall” es la capacidad de poder detectar correctamente la enfermedad entre los enfermos o lo que sería igual a: (Verdaderos positivos / Total Enfermos).

La Especificidad (“Especificity”), también conocida como la Tasa de Verdaderos Negativos, (“True Negative Rate”) o TN. Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuan bien puede el modelo detectar esa clase. Se calcula así:

$$\frac{VN}{(VN + FP)}$$

En el área de la salud se dice que la especificidad es la capacidad de poder identificar los casos de pacientes sanos entre todos los sanos (Verdaderos Negativos / Total Sanos).

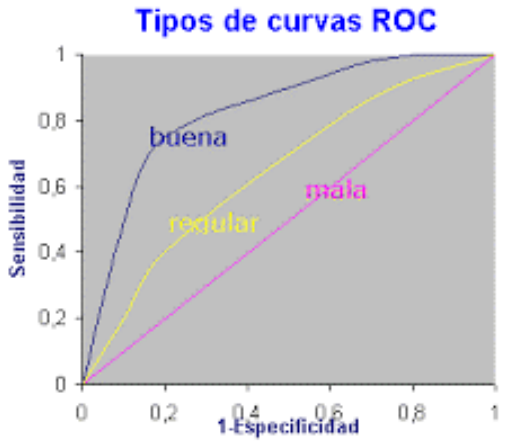
F1 Score: El F1 Score es una métrica que combina tanto la precisión como la sensibilidad. Se calcula como la media armónica entre la precisión y la sensibilidad. Esta métrica es útil cuando se tiene un conjunto de datos desequilibrado, en el cual la cantidad de muestras negativas es mucho mayor que la de muestras positivas.

$$F1 = \frac{2 * Precision * Sensibilidad}{Precision * Sensibilidad}$$

Curva ROC (Receiver Operating Characteristic)

La curva ROC es una representación gráfica de la capacidad de un modelo para distinguir entre clases positivas y negativas. Es la representación de la razón o proporción de verdaderos positivos (TPR = Razón de Verdaderos Positivos) frente a la razón o proporción de falsos positivos (FPR = Razón de Falsos Positivos) según se varía el umbral de discriminación. La curva ROC permite evaluar y comparar el desempeño de diferentes modelos de clasificación binaria. Cuanto más se acerque la curva al vértice superior izquierdo del gráfico, mejor será el desempeño del modelo.

Figura 10 . Curva ROC



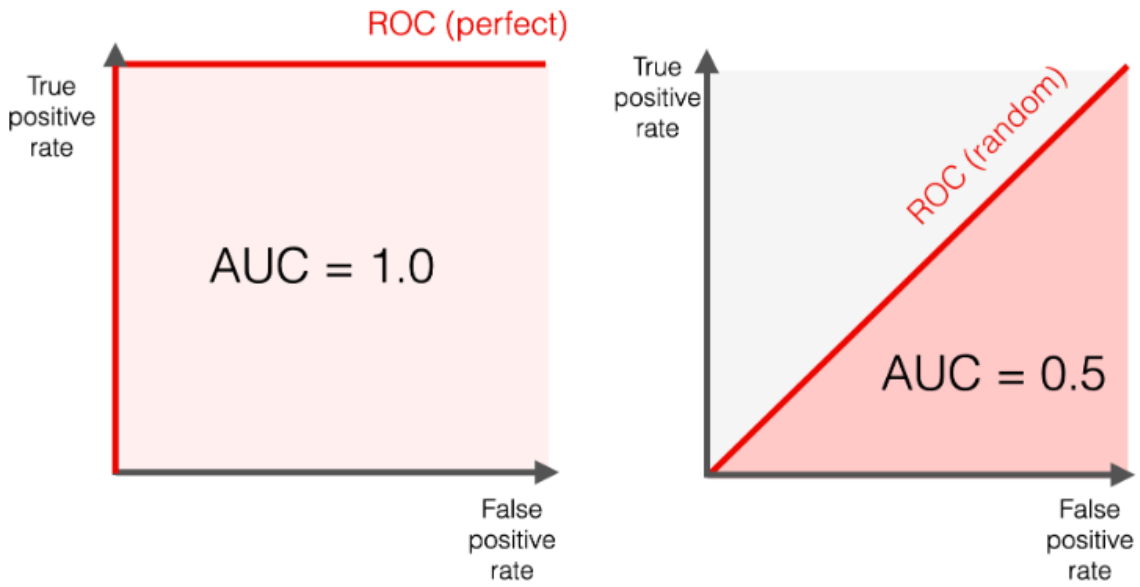
AUC: Área bajo la curva ROC

ROC-AUC es una métrica popular para evaluar el rendimiento de los modelos de clasificación binaria. Para calcularlo, se debe medir el área bajo la curva ROC, que muestra el desempeño del clasificador en diferentes umbrales de decisión. ROC-AUC puede variar de 0 a 1, una puntuación de 0,5 indica conjeturas aleatorias y una puntuación de 1 indica un desempeño perfecto.

Una puntuación ligeramente superior a 0,5 muestra que un modelo tiene al menos "algo" (aunque pequeño) poder predictivo. Como regla general, una puntuación AUC superior a 0,7 se considera buena, mientras que una puntuación superior a 0,9

se considera excelente. Sin embargo, la utilidad del modelo depende del problema específico y del caso de uso, no existe ningún estándar. Se debe interpretar la puntuación ROC-AUC en contexto, junto con otras métricas de calidad de clasificación.

Figura 11 - ROC-AUC



ESTADO DEL ARTE

Se han propuesto varios modelos predictivos en la literatura para la predicción de la cancelación. La eficiencia de cualquiera de estos modelos en gran medida está basada en la selección de los atributos del cliente (selección de características o variables) del conjunto de datos para la construcción del modelo. Para la elaboración de este proyecto, se recopilaron gran cantidad de artículos clasificándolos en función de las técnicas de minería utilizadas en ellos. Como se ve en la tabla, ordenadas en número de artículos en ese campo y la actualidad de la publicación:

Técnica Data Mining	No Publicaciones
Redes Neuronales	[5] [6] [7] [8] [9] [10]
Arboles de Decisión	[5] [7] [9] [10] [11] [12]
Regresión Logística	[8] [13] [14] [15]
Bosques Aleatorios	[10] [13] [15] [16]
SVM	[10] [12] [14] [17]

Tabla 1: Artículos recopilados por técnicas de minería

En Hu, X. (2020, abril 20). Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network, los autores presentan un enfoque híbrido para calcular la rotación de clientes mediante la combinación de árboles de decisión y redes neuronales. Luego se evaluó el método utilizando datos de almacenes y supermercados. Como en la gran mayoría de trabajos sobre cancelación, se requiere una gran numero de datos atractivos. Sin embargo, los autores no proporcionaron una comparación del con otros modelos. Los autores en Gunther, C.C., et al.: Modelling and predicting customer churn from an insurance company. Scand. Actuar. J. 2014(1), 58–71 (2014), analizan datos de seguros usando regresión logística. Esta investigación utiliza datos del sector de

seguros, pero aquí, los investigadores se beneficiaron de datos más detallados, que proporcionaron una vista de los datos mes a mes. Como resultado, los autores no vieron la necesidad de proporcionar una comparación entre el modelo y otros algoritmos de clasificación. En Bolance, C., Guillen, M., Padilla-Barreto, A.E.: Predicting probability of customer churn in insurance. In: Leon, R., Muñoz-Torres, M., Moneva, J. (eds.) MS 2016. LNBIP, vol. 254, pp. 82–91. Springer, Cham (2016). los autores presentan un medio para predecir la cancelación de clientes para el seguro de automóviles, comparan cuatro métodos utilizados para calcular la cancelación: árboles de decisión, redes neuronales, regresión logística y máquinas de vectores de soporte. En Sundarkumar, G.G., Ravi, V., Siddeshwar, V.: One-class support vector machine based undersampling: application to churn prediction and insurance fraud detection. In: 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–7. IEEE (2015), fueron empelados (5) métodos de clasificación para evaluar la eficiencia de este estudio, los cinco métodos fueron: arboles de decisión, SVM, regresión logística, redes neuronales.

OBJETIVO

El principal objetivo que se persigue con la implementación de este trabajo es el de desarrollar un modelo que permita predecir los clientes que cancelaran el seguro voluntario en los siguientes (**dos**) meses

Objetivos Específicos

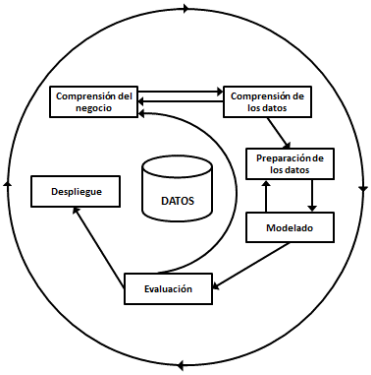
- Construir un repositorio de datos centralizado que incluya la mayor cantidad de información relevante que sea posible obtener acerca del cliente, este proceso es de vital importancia para poder estructurar la información que sustente las decisiones que se tomen en la implementación del modelo, adicionalmente será un insumo para futuros evoluciones del modelo.
- Comparar el desempeño de al menos cinco técnicas de aprendizaje automático tales como árboles de decisión, random forest, máquinas de soporte vectorial, naive bayes y XGBoost entre otros.
- Identificar aquellos clientes que tienen una probabilidad alta de cancelación, para iniciar una estrategia comercial de retención.

METODOLOGIA

La metodología CRISP-DM es ampliamente reconocida y utilizada en el campo de la minería de datos y el análisis de datos, proporcionando una estructura bien definida para llevar a cabo un proyecto en estos campos, garantizando que la investigación o el proyecto siga un proceso sistemático y bien organizado, lo que contribuye al rigor metodológico necesario para su desarrollo. Es una metodología altamente adaptable que puede aplicarse a una amplia variedad de problemas, además abarca todas las etapas del ciclo de vida de un proyecto garantizando que se cubran todas las fases necesarias para su éxito. Particularmente se centra en la comprensión de los objetivos del negocio y en la generación de valor a través del análisis de datos del negocio, lo cual es altamente esencial en el campo de la investigación académica, ya que ayuda a demostrar la relevancia y el impacto del trabajo en el mundo real. Finalmente, los proyectos que son desarrollados con esta metodología son ampliamente aceptados tanto en la comunidad académica como en el mundo empresarial, lo que puede aumentar la credibilidad del trabajo y su aplicabilidad práctica.

En el ámbito de la metodología CRISP-DM, se establece las diferentes etapas para el desarrollo de los objetivos propuesto en el presente trabajo. Como se observa en la Figura 12

Figura 12 - Metodología CRISP-DM



Fuente: CRISP-DM: Towards a standard process model for data mining (2000)

Comprensión del negocio

En Colombia, el sector asegurador ha tenido una acelerada transformación caracterizada por el surgimiento de nuevos canales de comercialización, un mayor uso de la tecnología y cambios en el comportamiento de los consumidores.

Figura 13 - Industria aseguradora año 2022

COMPORTAMIENTO DE LAS PRIMAS EMITIDAS
(miles de millones de pesos)

RAMOS	PRIMAS EMITIDAS						
	dic-20	dic-21	Δ	dic-22	Δ	Part.	Contr.
DAÑOS	10.972,6	13.015,6	18,6%	15.772,5	21,2%	33,3%	7,8%
PERSONAS	9.183,5	10.225,3	11,3%	12.085,9	18,2%	25,5%	5,3%
SEGURIDAD SOCIAL	7.474,9	8.918,9	19,3%	15.870,4	77,9%	33,5%	19,7%
SOAT	2.883,5	3.184,2	10,4%	3.584,3	12,6%	7,6%	1,1%
TOTAL INDUSTRIA	30.514,5	35.344,0	15,8%	47.313,2	33,9%	100%	33,9%

Fuente: Fasecolda, Estadísticas De La Industria Aseguradora Y De Capitalización. Diciembre-2022

En los últimos años, el sector ha tenido un buen desempeño. En 2022 la industria aseguradora emitió primas por un valor de $47.3 billones, el indicador de penetración de la industria llegó al 3% del PIB y el gasto anual real por habitante en seguros equivale a $611.505.

En el análisis del sector desde el lado de la demanda, la Encuesta de Demanda de Inclusión Financiera 2022 dejó ver que los seguros voluntarios que los colombianos más afirman tener son los exequiales (14,4%), seguido de los seguros de vida (12,9%), todo riesgo (5,3%), de accidentes personales (3,8%), para el hogar (3%) y de desempleo (2,5%). Adicionalmente, si se hace un perfil del colombiano que se asegura de manera voluntaria, se evidencia que los hombres son los de mayor participación, con 34% de los encuestados; mientras que en el caso de las mujeres es de 26,1%. Mientras que, por rango de ingresos entre más aumenta, más son las personas que se aseguran. Por regiones, en el Centro Oriente y el Pacífico

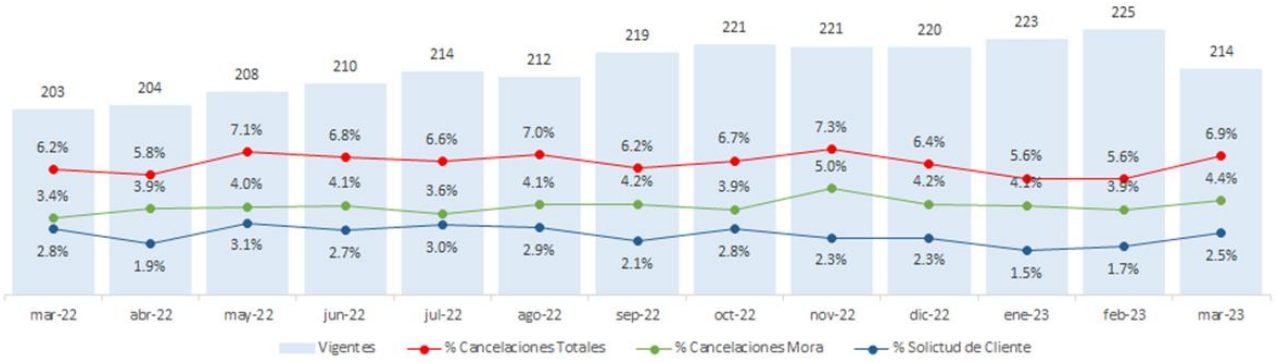
es donde más se aseguran, con 30,8% y 27,8% respectivamente. Le siguen Eje Cafetero (26,5%), Centro Sur (23,3%), Llano (19,7%) y Caribe (16,3%).

Por otro lado, para el año 2.022 el negocio de Seguros represento para la entidad un ingreso por $13.780 mm de pesos incrementando un 4% con respecto al año anterior, lo cual es un índice bajo si se tiene en cuenta que el IPC de este mismo año fue del 13,12%. Y si se compara con los ingresos obtenidos en el año 2,019 antes de la pandemia este valor se ha reducido en un 19%.

Tipo Seguro	2019	2020	2021	2022	%Crec
Libre Inversion	$ 5,519	$ 4,362	$ 2,882	$ 2,705	-6%
Cuentas	$ 7,182	$ 6,326	$ 5,974	$ 5,729	-4%
Libranzas	$ 1,678	$ 2,003	$ 2,211	$ 2,518	14%
Tarjetas	$ 847	$ 771	$ 621	$ 650	5%
Incentivos	$ 1,787	$ 1,764	$ 1,588	$ 2,186	38%
Total general	**$ 17,012**	**$ 15,229**	**$ 13,278**	**$ 13,789**	**4%**

Tabla 2: Ingresos por comisiones de seguros 2019 – 2022 - Fuente Banco AV Villas

Figura 14 - Cancelaciones históricas



Fuente: Banco AV Villas – Dirección de Bancaseguros

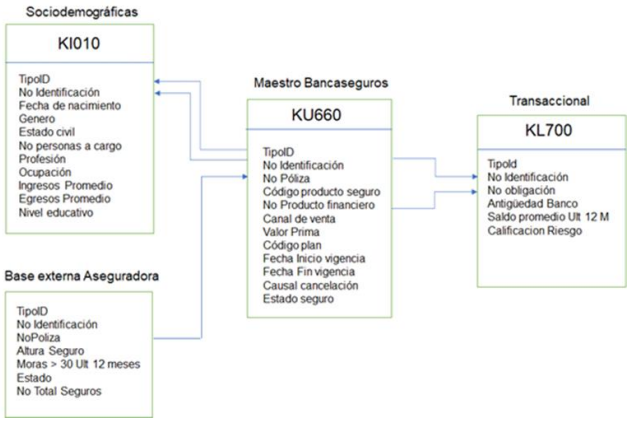
Particularmente cuando se analiza los datos de cancelación al cierre del año 2.022, se encontró que la tasa de cancelación de clientes es del 2.3% del total de clientes activos del periodo (220.000 clientes), esto son alrededor de 4.500 clientes al mes, Al cierre de marzo del presente año esta tasa está situada en 2.5% con un total de 214.000 clientes vigentes.

Compresión de los datos

Para este trabajo, se extrajo una base de clientes asegurados y con productos de cuentas de ahorro y corriente vigentes para el año 2.022, para estos clientes, la compañía vende siete tipos principales de seguros voluntarios: Fraude, Vida, Accidentes Personales, Exequial, Cáncer, Hogar y Desempleo. Además, la compañía también ofrece varios tipos de seguros adicionales. Los datos de los clientes activos al cierre de diciembre de 2.022 se extrajeron utilizando SQL server de la base de datos de la compañía. El período comprendido entre 01-10-2012 y el 30-11-2022 se utiliza para seguir el comportamiento del cliente e identificar si el cliente ha cancelado (terminado) o no (no terminado). Un cliente se define como activo si tiene al menos un tipo de cobertura en uno de los siete tipos de seguros principales o uno de los tipos de seguros adicionales. Un cliente se ha dado de baja (terminado, = 1) durante el período seleccionado, si todos los tipos de cobertura están cancelados y el cliente no tiene ningún tipo de cobertura activo dentro de ninguno de los tipos de seguros.

Para este objetivo se realizó la extracción de la información a través de ETL a los maestros de información de clientes [KI010], maestro de información de seguros [KU660] y el maestro transaccional [KL700], la información correspondiente a la aseguradora fue obtenida a través de un servicio web de intercambio de información cruzando por el número de identificación del cliente. El modelo de datos desarrollado se representa en la figura 15.

Figura 15 - Modelo de datos



Fuente: Banco AV Villas – Dirección Bancaseguros

El set de datos tiene está compuesto por (14) variables de las cuales (9) corresponden a variables numéricas y (5) categóricas, así:

Variable	Descripción	Tipo
EdadCli	Edad cliente	numérica
EstadoCivil	Estado civil cliente	categórica
TipoOcupacion	Actividad económica cliente	categórica
Genero	Genero cliente	categórica
NroPersonasACargo	No Hijos o personas dependientes	numérica
IngresoMensual	Salario mensual cliente	numérica
SegmentoCli	Segmento mercado cliente	categórica
NumSeguros	Total seguros adquiridos cliente	numérica
Vlr_Prim_Seg	Valor prima seguros adquiridos cliente	numérica
AltSeg	Altura seguros adquiridos cliente	numérica
Total_CtasAHO	Total cuentas de ahorro cliente	numérica
Sldctas_Cierre	Saldo cuentas cliente al cierre	numérica
ProCts_cierre	Promedio cuentas cliente al cierre	numérica
terminación	Indicador de cancelación	categórica

Tabla 3: Lista de variables conjunto de datos

Cabe mencionar que en la obtención del set de datos final se descartaron una serie de variables asociadas a la identificación del cliente y la póliza, tales como: # de Identificación cliente, # de póliza de seguro, Teléfono celular cliente, Dirección domicilio cliente. Se tomó la decisión de descartar estas variables ya que no se encontró una mayor relación con la variable objetivo que queremos predecir.

Preparación de los datos

En la mayoría de los problemas de la vida real, los datos son incompletos e inconsistentes debido a errores humanos o informáticos, a las deficiencias en la transmisión de datos o de diferentes fuentes de datos. Por lo tanto, la limpieza, la integración, la transformación, la reducción de datos y la discretización de los datos son principales tareas en su preprocesamiento (Han, Kamber y Pei, 2011).

Limpieza: Para evitar datos incompletos, ruidosos y/o inconsistentes, al seleccionar los datos se utilizan los siguientes criterios:

- Clientes sin información completa sobre alguna de las características extraídas fueron excluidas del set de datos para evitar datos faltantes. En este proceso se eliminaron 10 registros en su totalidad lo cual corresponde al 0.04% del total de los datos.

- Se excluyeron clientes con el rango de edad menor a 18 años, ya que a esta edad no es permitido adquirir un seguro.

Preprocesamiento**:** El primer enfoque es identificar los distintos valores existentes en las variables categóricas como ['EstadoCivil', 'TipoOcupacion', "Genero", y "SegmentoCli"] y sustituir cada uno de ellos por un número normalmente enteros

que inician en 0 y así convertirlos en un formato legible para la máquina y de esta manera los algoritmos de aprendizaje automático pueden decidir de una mejor manera cómo se deben gestionar esas etiquetas. En segunda instancia se decidió escalar las variables numéricas, tales como: ['EdadCli', 'NroPersonasACargo', 'NumSeguros', "AltSeg", "Total_CtasAHO"]

Modelamiento

En esta fase se lleva a cabo la generación de modelos de aprendizaje a partir de los datos suministrados desde la fase anterior. Se utilizaron siete algoritmos de aprendizaje : Regresión Logística, K-Vecinos más cercanos, Random Forest, Decision Tree, Naive Bayes, Máquinas de Soporte Vectorial - SVM y XGBoost.

Para estos algoritmos seleccionado se debe realizo una partición de los registros del dataset 70% para entrenamiento y 30% validación del modelo, Adicionalmente, se ha utilizado la técnica de validación cruzada la cual consiste en separar en fragmentos de entrenamiento y validación de manera iterativa, y su utilización para la obtención de las métricas con las que evaluaremos su desempeño.

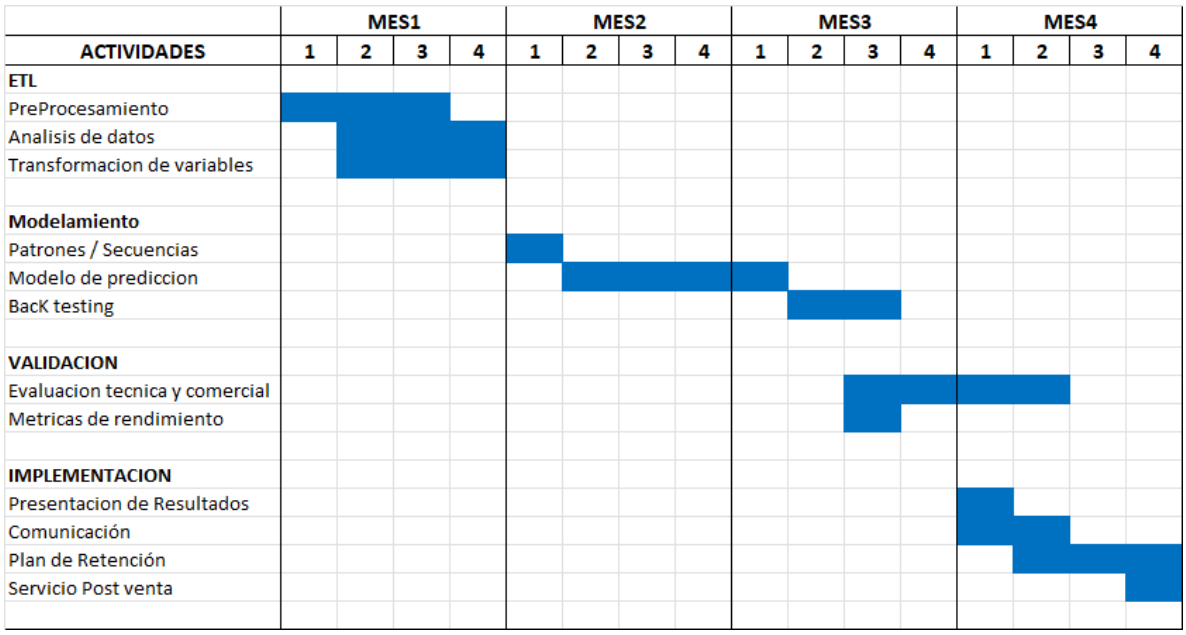
Evaluación de resultados: Si los modelos obtenidos cumplen con las expectativas del negocio, se procede a la explotación del modelo, Si no se cumple con esta expectativa, se evalúa si se procede a iterar nuevamente sobre los pasos anteriores con el objetivo de encontrar nuevos resultados.

Despliegue: En esta última fase se definen las estrategias para la implementación, monitoreo y mantenimiento de modelo seleccionado.

CRONOGRAMA DE ACTIVIDADES

Se determinó que el proyecto tendrá una duración de cuatro (4) meses iniciando desde enero de 2.023. Las actividades que se desarrollaran para implementar el modelo de predicción de cancelación están descritas a alto nivel en la siguiente figura.

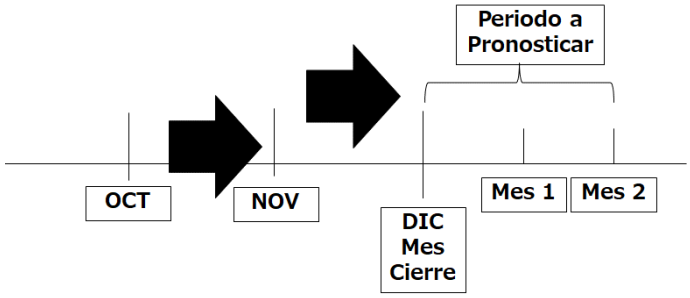
Figura 16 - Cronograma de actividades



RESULTADOS

Para este proyecto inicialmente se incluyeron dentro del análisis 132.000 clientes que habían adquirido el seguro voluntario a lo largo del año 2.022 y que para el cierre de diciembre de este año el producto de seguro se encontraba aún vigente, posteriormente se acotó la muestra de estudio a 22.238 registros los cuales describen a aquellos clientes que adquirieron específicamente el seguro entre los meses de octubre y noviembre que se encontraban activos al cierre de diciembre y que posteriormente cancelaron el producto en los dos siguientes meses del año. En la figura se ilustra la estrategia seguida con los datos.

Figura 17-Estrategia ETL



A continuación, se realiza una operación que es común para todos los modelos de aprendizaje supervisado, que es la división del conjunto de datos en dos partes: una parte Train, de entrenamiento, que corresponderá a la mayor parte del dataset y que se usará para entrenar el modelo y una parte Test, de menor tamaño, sobre la que evaluaremos nuestro modelo entrenado. La parte Test se deja reservada del entorno de trabajo y únicamente se utilizará cuando se haga la correspondiente evaluación final del modelo.

Análisis descriptivo de los Datos

El set de datos (22,238 registros) desde el punto de vista descriptivo tiene las siguientes características para las variables numéricas:

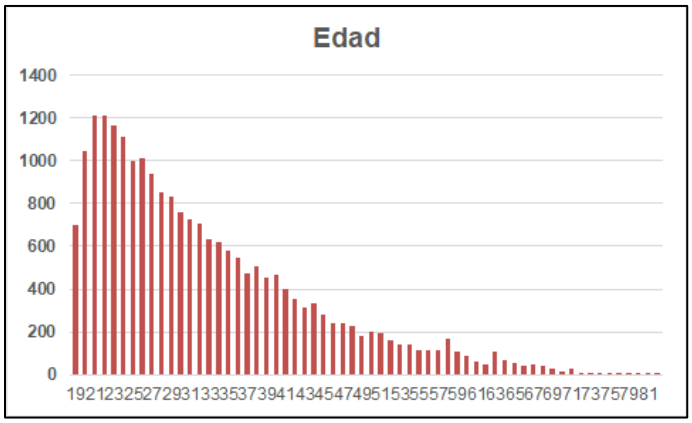
Figura 18 - Análisis descriptivo datos numéricos

	Edad Cliente	No Personas a cargo	IngresoMensual	Num Seguros	Valor Prima Seguro	Altura seguro	Total Productos Financiero	Saldo productos	Promedio productos
mean	32	0.1	1,556,256	1.1	18,372	1.5	2	769,054	496,009
std	11	0.4	2,733,576	0.3	13,957	0.8	1	4,048,621	3,613,126
min	19	-	1,160,000	1.0	11,700	-	1	- 0	-
0.25	24	-	1,160,000	1.0	14,600	0.7	1	1,429	18,194
0.50	30	-	1,160,000	1.0	17,050	1.4	1	61,020	72,452
0.75	39	-	1,200,000	1.0	18,600	2.2	2	666,072	178,211
max	84	4.0	150,000,000	4.0	1,868,900	3.0	16	184,689,900	175,024,500

Fuente: Banco AV Villas – Dirección Bancaseguros

La edad promedio de los clientes es de 32 años, con un valor mínimo de 19 y máximo de 84 años, es importante anotar que el 75% de los datos representan a una población con 39 años o menos reflejando así la marcada juventud de la población.

Figura 19 - Variable edad



Fuente: Banco AV Villas – Dirección Bancaseguros

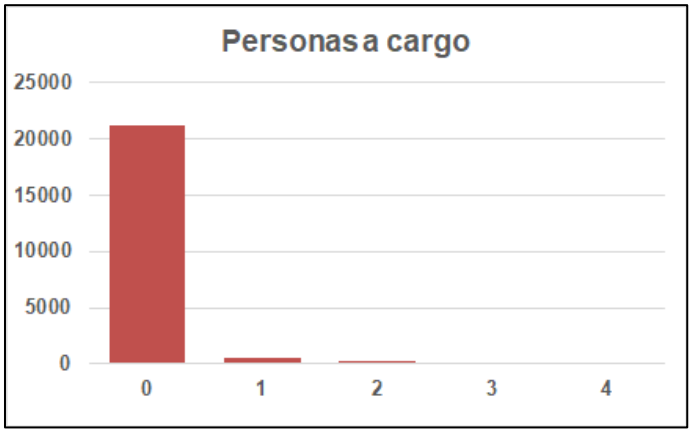
En promedio los clientes que adquirieron el seguro tienen 1.5 meses con el producto, se puede observar una distribución homogénea en cada uno de los meses de emisión, lo que denota un mercado estable en términos de adquisición del producto.

Figura 20 - Altura del seguro

Fuente: Banco AV Villas – Dirección Bancaseguros

Al ser una población relativamente joven, 50% de los clientes tienen 30 años o menos, puede ser razonable encontrar que solamente el 5% de ellos tienen personas a cargo (hijos o dependientes).

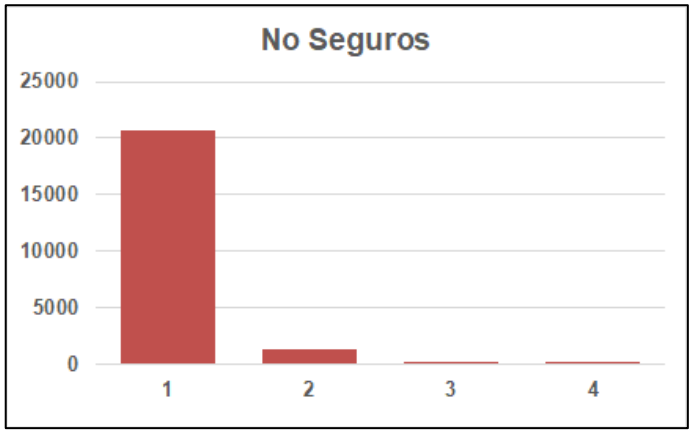
Figura 21 - Personas a cargo



Fuente: Banco AV Villas – Dirección Bancaseguros

El 93% de los clientes adquirieron únicamente un seguro, lo cual está relacionado a la estrategia comercial definida por la institución para el ofrecimiento de este producto. Para el 7% de los clientes, se puede asumir que poseen ya una cultura de aseguramiento en razón a que adquieren más de un producto de seguros al momento de acercarse a la institución.

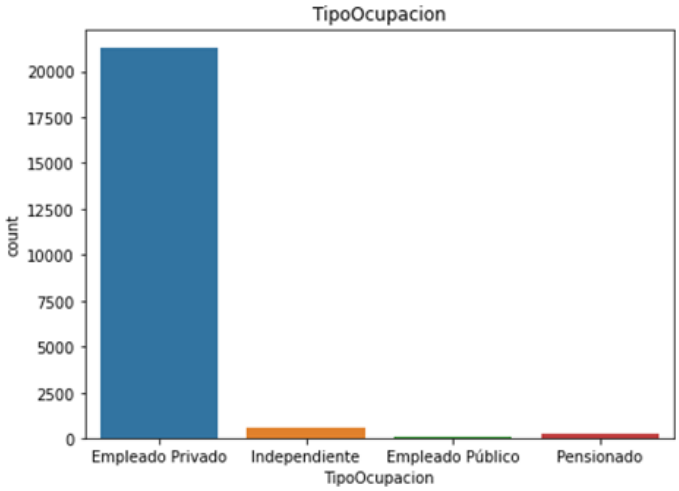
Figura 22 - Numero seguros posee el cliente



Fuente: Banco AV Villas – Dirección Bancaseguros

Para las variables categóricas, iniciamos el análisis con la variable “actividad económica”, para esta población se encontró que tienen como principal actividad “Empleado” con el 95.5% de participación, esto refleja fidedignamente el segmento de clientes al cual la institución le ofrece sus productos y servicios y a cual actividad económica prefiere en sus clientes, cabe la pena recordar que los seguros voluntarios no se ofrecen de manera individual, sino que siempre se ofrecen asociados a un producto financiero.

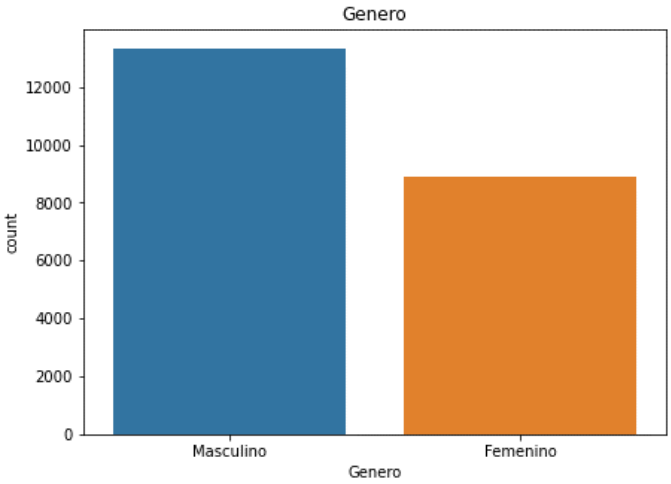
Figura 23 - Tipo de ocupación



Fuente: Banco AV Villas – Dirección Bancaseguros

Hay una diferencia marcada en la composición de la base cuando se habla en términos de la variable “Género”, ya que la participación de hombres es del 59.9% y de mujeres es del 40.1%.

Figura 24 - Genero

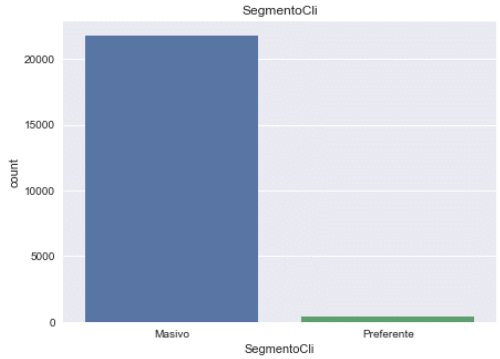


Fuente: Banco AV Villas – Dirección Bancaseguros

De acuerdo con las políticas establecidas de colocación de productos, en la institución hay una marcada tendencia de ofrecimiento a los clientes de segmento

masivo, los cuales son catalogados de acuerdo con su nivel de ingreso. En ese sentido, los clientes que tienen ingresos inferiores a seis (6) salarios mínimos($6.000.000) son catalogados en el segmento masivo.

Figura 25 - Segmento

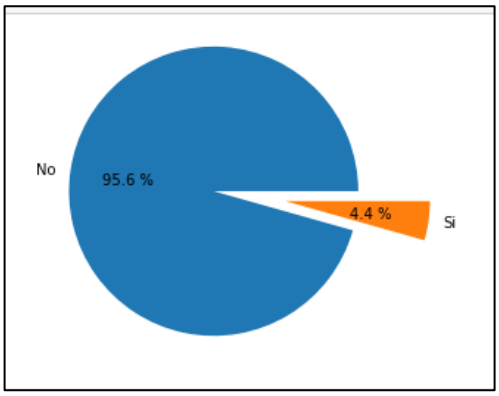


Fuente: Banco AV Villas – Dirección Bancaseguros

Variable Objetivo

Para los clientes que se encontraban activos al cierre de diciembre de 2.022 se presentaron 971 cancelaciones que se dieron voluntariamente, lo que indica una tasa de Churn del 4.4 %. Debido a que el porcentaje de seguros cancelados no es igual a los activos, el conjunto de datos está desbalanceado, esto es, un conjunto de datos donde el número de observaciones pertenecientes a un grupo o clase es significativamente mayor que las pertenecientes a las otras clases.

Figura 26 - Variable objetivo



Fuente: Banco AV Villas – Dirección Bancaseguros

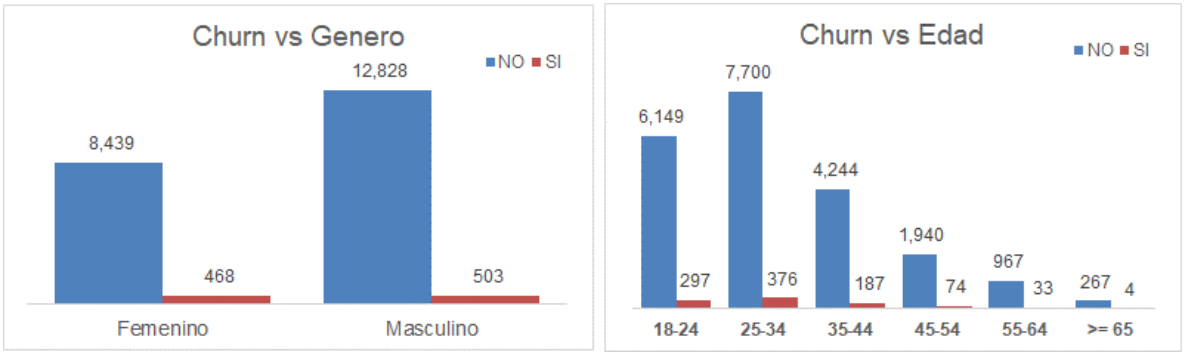
Tener un conjunto de datos desbalanceados representa un problema, ya que esto puede afectar la calidad del modelo y su capacidad para predecir correctamente. Es por esto, por lo que se utilizaron varias estrategias para tratar con estos tipos de datos:

- Ajuste de Parámetros del modelo: Se ajustaron parámetros del propio algoritmo para intentar equilibrar a la clase minoritaria penalizando a la clase mayoritaria durante el entrenamiento. Por ejemplo, ajuste de peso en árboles, y en regresión logística a través del parámetro class_weight= “balanced”. No todos los algoritmos tienen estas posibilidades.
- Utilización de las métricas de desempeño apropiadas, como el Recall y el F1 Score.
- No se crearon muestras sintéticas utilizando la técnica SMOTE [18], ya que al crear nuevos registros para balancear un poco mejor las clases, podemos alterar la distribución natural de esa clase (clase minoritaria) y confundir al modelo en su clasificación a través del “sesgo” que pueden tener los datos originales.

Cuando se analiza la variable objetivo frente a las variables categóricas, encontramos:

La tasa de cancelación en mujeres es un 30% más alta (5.55%) respecto a la de los hombres (3.92%). El grupo de personas que tienen 45 años o menos de edad en promedio tienen una tasa de cancelación superior en un 60% al resto de la población, esto es (4.71%) frente a un (2.91%) respectivamente.

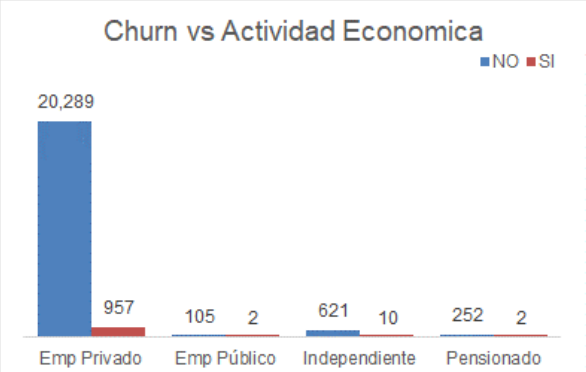
Figura 27 - Genero y edad vs Tasa de cancelación

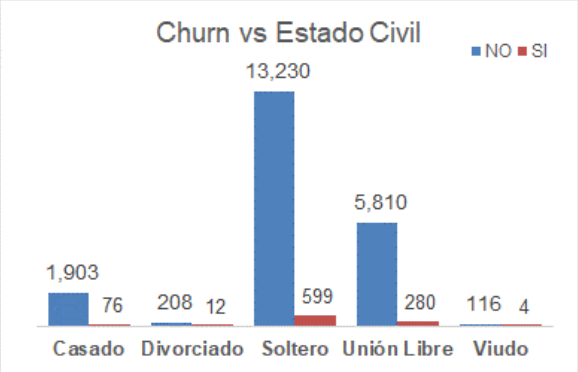


Fuente: Banco AV Villas – Dirección Bancaseguros

La tasa de cancelación en empleados privados es 2.2 veces mayor (4.72%) respecto al promedio de las demás actividades económicas (1.44%), característica importante, ya que como se describió anteriormente cerca del 96% del set de datos tienen este tipo de actividad económica. El grupo de personas estado civil casadas tienen una tasa de cancelación del (3.99%) frente a un (4.67%) de las personas con estado civil soltero, esto es un 14% menor. Ver figura 18

Figura 28 - Actividad económica y estado civil vs Tasa de cancelación

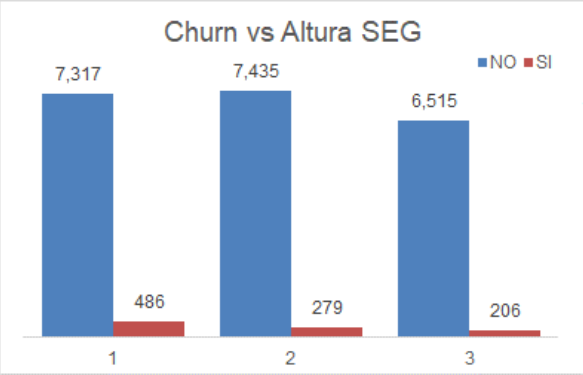


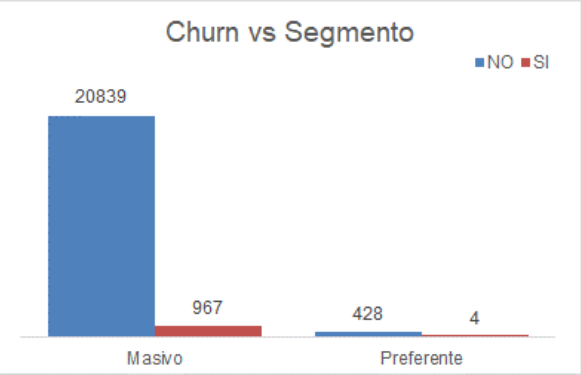


Fuente: Banco AV Villas – Dirección Bancaseguros

La tasa de cancelación de aquellos clientes que tienen un (1) mes de altura en el seguro (6.64%) dobla al resto de la categoría, esto es preocupante desde el punto de vista comercial porque puede indicar algún grado de malas ventas o inducción al cliente a tomar el producto sin la información necesaria. La tasa de cancelación que presenta los clientes preferentes (0.93%) puede demostrar que la permanencia en el seguro es sensible al precio, es decir, a mayor nivel de ingresos mayor permanencia del seguro.

Figura 29 - Altura seguro y segmento vs Tasa de cancelación

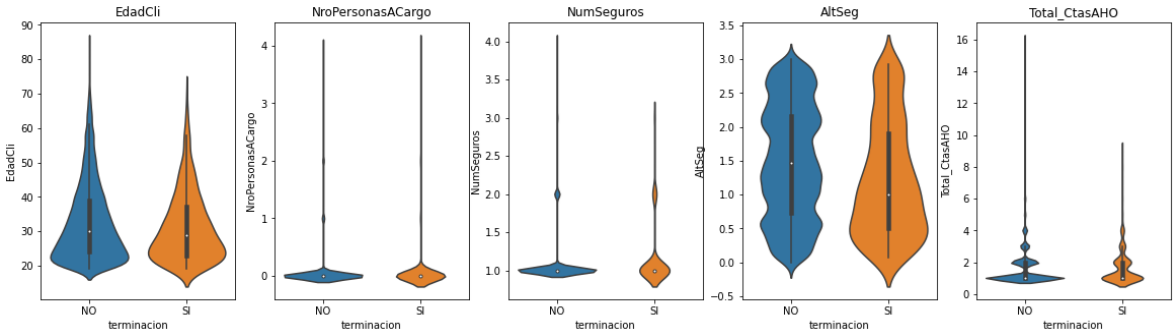




Fuente: Banco AV Villas – Dirección Bancaseguros

Cuando se analiza la variable objetivo frente a las variables numéricas, encontramos:

Figura 30 - Variables numéricas vs Tasa de cancelación



Fuente: Banco AV Villas – Dirección Bancaseguros

Se puede observar un comportamiento similar del Churn en las variables numéricas con excepción al Numero de seguros por cliente, en donde se aprecia un grado de cancelación mayor para los clientes que poseen (2) seguros y también para los clientes que son relativamente jóvenes en la adquisición del seguro, ya que se aprecia una cancelación mayor en los clientes que tienen Altura menor a un mes de expedido.

Entrenamiento de los Modelos

Como la intención primordial del modelo es poder predecir el mayor número de clientes que van a cancelar el seguro voluntario, se aplicaron siete modelos de aprendizaje automático a los datos de entrenamiento, realizando a su vez validación cruzada utilizando la técnica k-folds *(cv = 5 pliegues)* y determinado la métrica ROC-AUC como medida para determinar con cuál de los modelos aplicados se obtiene una mayor capacidad predictiva. Y es aquí, desde el punto de vista de la interpretación, la métrica ROC-AUC es la medida más útil porque nos determina qué tan bien se clasificaron las predicciones del modelo, mostrándonos cuál es la

probabilidad de que una instancia positiva elegida al azar tenga una clasificación más alta que una instancia negativa elegida al azar. Cuando se tienen datos desbalanceados, la métrica *accuracy* puede ser engañosa, Incluso un modelo que predice siempre la clase mayoritaria puede tener un alto *accuracy* debido a la distribución desigual de las clases. ROC-AUC es menos sensible a desbalances en los datos y proporciona una evaluación más robusta del rendimiento del modelo.

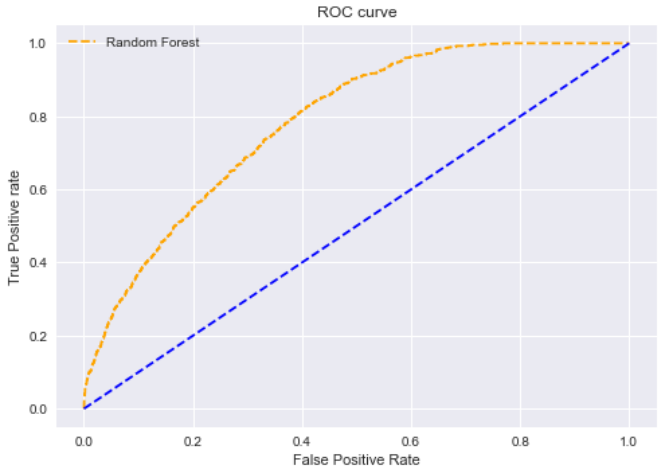
En algunos casos, los errores de clasificación en una de las clases pueden ser más costosos o críticos que en la otra clase. Por ejemplo, en el desarrollo de este trabajo, predecir incorrectamente una persona que cancele su seguro puede ser más grave que predecir incorrectamente que un cliente no lo cancele. ROC-AUC considera la tasa de verdaderos positivos frente a la tasa de falsos positivos en todo el rango de umbrales de decisión, lo que puede ayudar a evaluar cómo el modelo maneja diferentes tasas de error en las clases.

Figura 31 - Comparativo score ROC-AUC

MODELO	SCORE (ROC_AUC)
Random Forest	**0,6552**
Regresión Logística	0,6401
Decision Tree	0,5118
Naive Byes	0,6074
KNN	0,5607
SVM	0,4915
XGBoost	0,6403

Como se puede observar en la tabla, la puntuación de la métrica ROC-AUC del modelo de Random Forest es la más alta alcanzada entre los siete modelos escogidos, la curva ROC de este modelo se observa en la figura, la cual permite intuir que el modelo escogido tiene un razonable desempeño de predicción.

Figura 32 - Curva ROC - Modelo Random Forest



En este punto, llevar a cabo una validación cruzada de 5 iteraciones genera cinco modelos, cinco fuentes de datos para entrenar los modelos, cinco fuentes de datos para evaluar los modelos y cinco evaluaciones, una para cada modelo. Entonces se genera una métrica de desempeño del modelo para cada evaluación y finalmente para obtener la medición del desempeño general del modelo se calcula la media de las cinco métricas ROC-AUC.

Los hiperparametros definidos para el modelo de Random Forest son los siguientes: *(Class_weight = 'balanced', criterion =' entropy', max_depth = 8, n_estimators=200, n_jobs = -1)*, en donde:

n_estimators = Número de árboles en el bosque

max_depth = Número máximo de niveles en cada árbol de decisión

Adicionalmente a manera de validación del desempeño de los modelos ejecutados se utilizó la librería [naiveautoML] en Python [18], la cual permite encontrar de manera simultánea cual podría ser el mejor algoritmo de preprocesamiento y aprendizaje (pipelines) con el mejor desempeño y rendimiento general para un determinado conjunto de datos, en este caso el conjunto de datos

de entrenamiento de seguros voluntarios utilizado previamente. Al ejecutar este recurso se validó que NO se iba encontrar un pipeline que tuviera un desempeño superior al desempeño previamente encontrado con el modelo de Random Forest aplicado.

Figura 33 - Visualización librería NaiveautoML

```
32it [13:54, 26.07s/it]
Progress for parameter turning:
100%|██████████| 10/10 [00:14<00:00,  1.48s/it]
Pipeline(steps=[('impute_and_binarize',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('imputer',
                                                                   SimpleImputer(strategy='median'))]),
                                                  ['EdadCli',
                                                   'NroPersonasACargo',
                                                   'IngresoMensual',
                                                   'NumSeguros', 'Vlr_Prim_Seg',
                                                   'AltSeg', 'Total_CtasAHO',
                                                   'SldCts_cierre',
                                                   'PrmCts_cierre']),
                                                 ('cat',
                                                  Pipeline(steps=[('imputer',
                                                                   SimpleImputer(strategy='most_frequent')),
                                                                  ('binarizer',
                                                                   OneHotEncoder(handle_unknown='ignore',
                                                                                 sparse=False))]),
                                                  ['EstadoCivil',
                                                   'TipoOcupacion', 'Genero',
                                                   'SegmentoCli'])])),
                ('data-pre-processor', Normalizer()),
                ('learner', LinearDiscriminantAnalysis())])
```

	score_internal	scores	new_best
0	0.5818	{'roc_auc': 0.5818}	True
1	0.5105	{'roc_auc': 0.5105}	False
2	0.5798	{'roc_auc': 0.5798}	False
3	0.5571	{'roc_auc': 0.5571}	False
4	NaN	{'roc_auc': nan}	False
5	0.5326	{'roc_auc': 0.5326}	False
6	0.6241	{'roc_auc': 0.6241}	True
7	NaN	{'roc_auc': nan}	False
8	NaN	{'roc_auc': nan}	False
9	NaN	{'roc_auc': nan}	False
10	NaN	{'roc_auc': nan}	False
11	NaN	{'roc_auc': nan}	False
12	0.5109	{'roc_auc': 0.5109}	False
13	NaN	{'roc_auc': nan}	False
14	0.6215	{'roc_auc': 0.6215}	False
15	0.6209	{'roc_auc': 0.6209}	False
16	0.6280	{'roc_auc': 0.628}	True
17	0.6236	{'roc_auc': 0.6236}	False
18	0.6244	{'roc_auc': 0.6244}	False
19	0.6232	{'roc_auc': 0.6232}	False
20	0.6234	{'roc_auc': 0.6234}	False
21	0.6241	{'roc_auc': 0.6241}	False
22	0.6272	{'roc_auc': 0.6272}	False
23	NaN	{'roc_auc': nan}	False
24	0.5003	{'roc_auc': 0.5003}	False
25	NaN	{'roc_auc': nan}	False

Encontrado el modelo y todavía en fase entrenamiento, se procede a calcular las probabilidades de las instancias positivas, es decir [Cancelar el Seguro], obteniendo una matriz de 2 columnas, con las siguientes características:

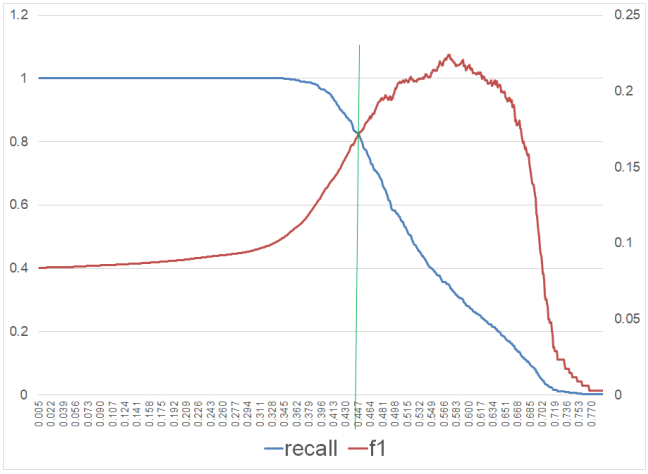
Figura 34 - Matriz entrenada de predicciones

```
array([[0.51221366, 0.48778634],
       [0.57173098, 0.42826902],
       [0.44916726, 0.55083274],
       ...,
       [0.51352619, 0.48647381],
       [0.47345996, 0.52654004],
       [0.64754164, 0.35245836]])
```

Dado que para este proyecto lo que más nos interesa es poder predecir la mayor cantidad de registros positivos , es decir, la mayor cantidad de personas que van a cancelar el seguro voluntario, debemos encontrar un umbral (Threshold) en la curva ROC en donde esto se maximice. Así que se procede a graficar el valor de Recall y

de F1 Score para cada una de las predicciones positivas entrenadas en el modelo. El umbral que cumple con las condiciones indicadas es el valor 0.441, en este punto el Recall es 0.852 y F1-Score es 0.165 como se observa en la figura.

Figura 35 - Gráfica Umbral optimo

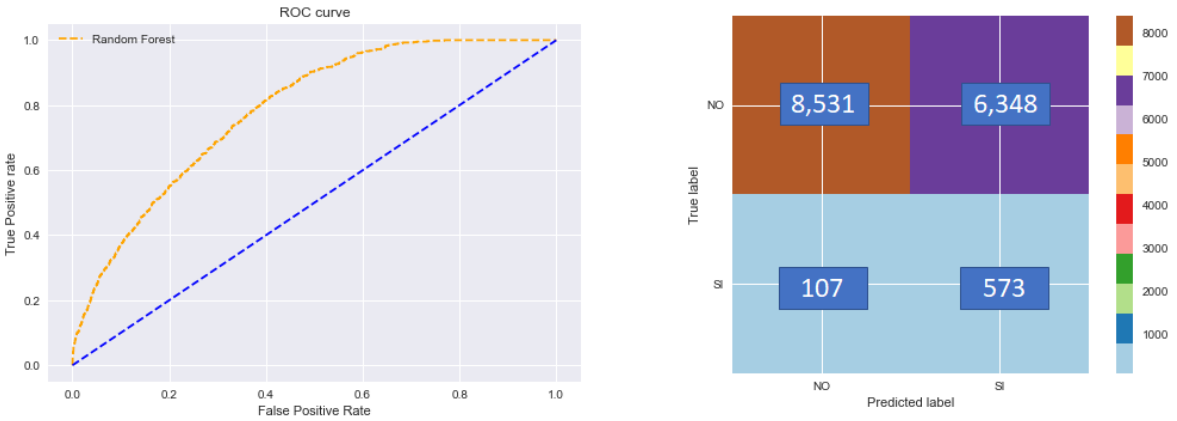


En el contexto del negocio, el objetivo de escoger este umbral es calificar a los clientes que probablemente cancelaran el seguro, para luego pasar esta lista de clientes de alto probabilidad de cancelación a un equipo del centro de llamadas para que se comunique con ellos. Es posible que el centro de llamadas no pueda comunicarse con todos ellos, pero fácilmente pueden llegar a un par de cientos, cada cliente que se retenga hará que el esfuerzo valga la pena. En este escenario, el costo de los falsos positivos es bajo (solo una gestión rápida que no resulta en cancelación), pero el valor de los verdaderos positivos es alto (ingresos inmediatos). En este caso, probablemente optimizaría para la cancelación, querer asegurarse de llegar a todos los clientes que potencialmente cancelaran. El único límite es la cantidad de personas que en el centro de llamadas puede contactar semanalmente. En este caso, se puede establecer un umbral de decisión más bajo. Es posible que el modelo tenga poca precisión, pero esto no es gran cosa siempre que se alcance

los objetivos comerciales y se realice una cierta cantidad de retenciones en el periodo.

La curva ROC y la matriz de confusión para este umbral son los siguientes:

Figura 36 - Curva ROC y Matriz de confusión UMBRAL



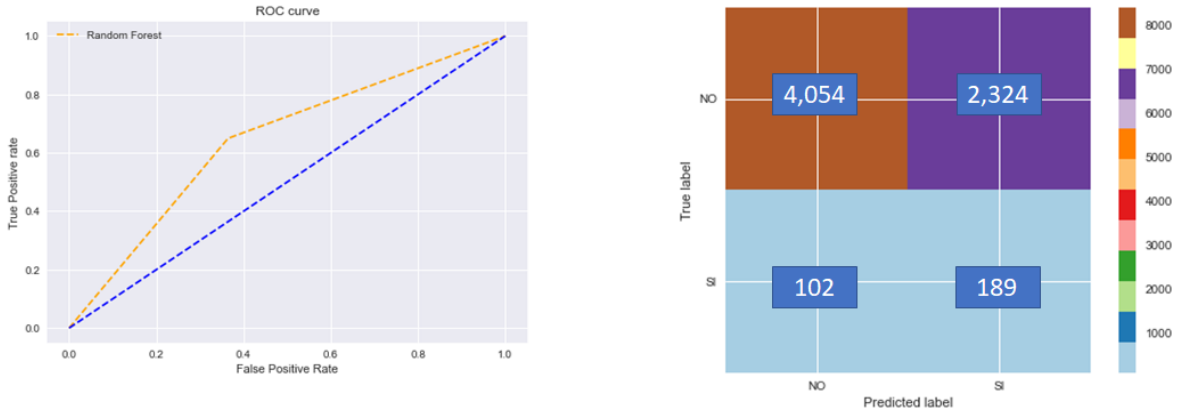
Con este modelo se identificaron correctamente 573 clientes, quedaron 6.348 falsos positivos y 107 clientes como falsos negativos. El RECALL para esta matriz es de 0.842

Predicciones

Finalmente, se aplica el modelo de Random Forest sobre los datos de TEST (6.669 registros), obteniendo los siguientes resultados:

[roc_auc] = **0.6425**, Métrica similar a la obtenida en el proceso de entrenamiento del modelo y que nos permite confirmar que el modelo tiene una capacidad de discriminación mejor que la aleatoria (alrededor del 15% adicional).

Figura 37 - Curva ROC y matriz de confusión para los datos TEST



Con este modelo se identificaron correctamente 189 clientes, quedaron 2.324 falsos positivos y 102 clientes como falsos negativos. El RECALL para esta matriz es de 0.6495,

A partir de estos resultados y con el ánimo de llevar a la práctica este modelo y utilizando los datos de Testeo como prueba piloto, para cada 'ID de cliente' se puede obtener una variable adicional relacionada al puntaje de propensión que resalte la probabilidad de que este cliente realice la acción de cancelación. A partir de allí, podemos definir una serie de grupos de clientes con diferentes grados de propensión a los cuales se les puede realizar diferentes acciones comerciales o estratégicas dependiendo de este orden de prioridad.

Grupo Cancelacion	Probabilidad cancelar	No. Clientes	%Part
Grupo1 - Altisima CAN	**>= 60.00%**	413	6.2%
Grupo2 - Alta CAN	**50.00 - 59.99%**	1,268	19.0%
Grupo3 - Media probabilidad CAN	**40.00 - 49.99%**	1,386	20.8%
Grupo4 - Baja probabilidad CAN	**30.00 - 39.99%**	2,652	39.8%
Grupo5 - Ninguna CAN	**<= 29.99%**	950	14.2%
Total general		6,669	100.0%

Tabla 4: Grupos de clientes calificados por probabilidad de cancelar

A manera de ejemplo, es primera prioridad atender a los clientes del Grupo No.1 dada su alta probabilidad de cancelación. Por ejemplo, se pueden disponer de comunicaciones a través del Contact Center para llegar al cliente e indagar su percepción actual del producto, realizar una reventa de las características principales o comunicarle las ventajas y los beneficios asociados al producto que de pronto no ha percibido o utilizado. Así mismo, se puede decidir no realizar acción alguna con los clientes del Grupo No. 5 dada su baja probabilidad de cancelación, simplemente la acción para estos clientes seria monitorear su evolución y revisar si cambian sus condiciones que determinen si se mueven o no entre los grupos determinados.

CONCLUSIONES

Como se ha podido corroborar a lo largo del trabajo realizado, la tasa de cancelación de clientes (Churn Rate) es una métrica fundamental a tener en cuenta al momento de construir la estrategia empresarial de la compañía, especialmente en el sector financiero donde se abordan mercados de segmento masivo, Si de la misma forma como se originan o se ganan clientes no existe una genuina preocupación por mantenerlos, llegara el momento o ciclo económico en el que las nuevas colocaciones o entradas de clientes sean más bajas y la cancelación se incremente.

Todos los modelos adecuados para problemas de clasificación binaria son aplicables en la industria de Bancaseguros para predecir la tasa de churn de los clientes. La literatura buscada para este trabajo no contiene restricciones ni limitaciones para la aplicabilidad de estos modelos en esta industria o un conjunto de datos específico relacionado. Todos los modelos son adaptables y aplicables independientemente del dominio de datos utilizado.

Este trabajo evidencia la premisa de cuando la métrica de “precisión” de un modelo no es suficiente para determinar su desempeño, porque los datos desbalanceados pueden generar una falsa idea de un gran desempeño al evidenciar alta “precisión” pero bajo “Recall”. Además de la “precisión”, se deben considerar otras métricas, para contrastar los resultados de desempeño en comparación con las metas y objetivos comerciales.

Como oportunidades de mejoramiento al modelo debemos considerar la inclusión de nuevas variables relacionadas, por ejemplo, el registro de utilización de los servicios adicionales que prestan actualmente los seguros y las respuestas dadas por el cliente a la llamada de bienvenida que se realiza solamente algunos días después de que él toma el seguro, como herramienta de evaluación de la venta realizada. En este trabajo se utilizaron todas las variables extractadas y disponibles para predecir la posibilidad de cancelación del cliente. En futuros trabajos, se debe

investigar si el entrenamiento de los modelos funcionaría mejor cuando solo se utilizan las variables más influyentes.

En cuanto a futuras líneas de trabajo de trabajo se complementaría el análisis predictivo con la generación de un tablero de control que permita tener una visual rápida y clara de los clientes potenciales a cancelar el producto, facilitando la generación de estrategias de fidelización y la elaboración de campanas proactivas que le permitan a la empresa disminuir la cancelación y mejorar su relación con los clientes.

REFERENCIAS BIBLIOGRAFICAS

[1] Yu, S.B., Cao, J. and Kung, Z.W. (2012) A Review of Customer Churn Problem Research. Computer Integrated Manufacturing Systems

[2] Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron, O'Reilly Media, Inc. O'Reilly Media, Inc.

[3] Lior Rokach and Oded Maimon (2008). Data mining with decision trees: theory and applications. World Scientific.

[4] Fix, E.; Hodges, J.L. (1989). (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951) ». International Statistical Review / Revue Internationale de Statistique 57 (3): 233-238

[5] X. Hu, "A Data Mining Approach for Retailing Bank Customer Attrition Analysis," Applied Intelligence, vol. 22, 2005, pp. 47–60, Springer.

[6] H. S. Song, J. K. Kim, Y. B. Cho and S. H. Kim, "A Personalized Defection Detection and Prevention Procedure based on the SelfOrganizing Map and Association Rule Mining: Applied to Online Game Site," Artificial Intelligence Review, vol. 21, 2004, pp. 161–184.

[7] Y. M. Zhang, J. Y. Qi, H. Y. Shu, and J. T. Cao, "A Hybrid KNN-LR Classifier and its Application in Customer Churn Prediction," Proc. the IEEE International Conference on Systems, Man and Cybernetics, Oct. 2007, pp. 3265–3269.

[8] G. Song, D. Yang, L. Wu, T. Wang, Sh. Tang, "A Mixed Process Neural Network and its Application to Churn Prediction in Mobile Communications," Proc. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006.

[9] Shin-Yuan Hung, David C. Yen and Hsiu-Yu Wang, "Applying data mining to telecom churn management," Expert Systems with Applications, vol. 31, 2006, pp. 515–524.

[10] J. Zhaoa and Xing-Hua Dang, "Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example," Proc. 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008 (WiCOM'08), Oct. 2008, pp. 1-4.

[11] Bong-Horng Chu, Ming-Shian Tsai and Cheng-Seen Ho, "Toward a hybrid data mining model for customer retention," Knowledge-Based Systems, vol. 20, 2007, pp. 703–718.

[12] K. Coussement, Dirk Van den Poel, “Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers,” Expert Systems with Applications, vol. 36, 2009, pp. 6127–6134.

[13] K. Coussement and Dirk Van den Poel, “Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,” Expert Systems with Applications, vol. 34, 2008, pp. 313–327.

[14] S. Lessmann and S. Voß, “A reference model for customer-centric data mining with support vector machines,” European Journal of Operational Research, vol. 199 (2), Dec. 2009, pp. 520-530.

[15] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” Expert Systems with Applications, vol. 36, 2009, 4626–4636.

[16] W. Buckinx and D. Van den Poel, “Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting,” European Journal of Operational Research, vol. 164, 2005, pp. 252–268.

[17] Y. Xie and X. Li, “Churn Prediction with Linear Discriminant Boosting Algorithm,” Proc. the Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 2008.

[18] Mohr, F., Wever, M. Naive automated machine learning. Mach Learn 112, 1131–1170 (2023). https://doi.org/10.1007/s10994-022-06200-0

BIBLIOGRAFIA

- Dawkins, P.M. and Reichheld, F.F. (1990) Customer retention as a competitive weapon. Directors & Board Summer, 42-7.
- Scriney, M., Nie, D., & Roantree, M. (2020). Predicting Customer Churn for Insurance Data. En M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Big Data Analytics and Knowledge Discovery* (pp. 256-265). Springer International Publishing. https://doi.org/10.1007/978-3-030-59065-9_21
- Morik and K¨opcke (2014, septiembre 12). Analysing Customer Churn in Insurance Data – A Case Study. Knowledge Discovery in Databases: PKDD 2004 Springer, pp. 325-336.
- Siemes, T. (2016, noviembre 16). Churn prediction models tested and evaluated in the Dutch indemnity industry [Tesis de pregrado]. Open University of the Netherlands Faculty Management, Science and Technology, Netherlands
- Kingawa, E. D., & Hailu, T. T. (2022). Customer Churn Prediction Using Machine Learning Techniques: The case of Lion Insurance. Asian Journal of Basic Science & Research, 04(04), 60-73. https://doi.org/10.38177/AJBSR.2022.4407
- Sahand Khakabi, Mohammad R. Gholamian, and Morteza Namvar. Data Mining Applications in Customer Churn Management. 2010 International Conference on Intelligent Systems, Modelling and Simulation, January 2010
- Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. Scandinavian Actuarial Journal, 2014(1), 58-71.
- Soeini, R. A., & Rodpysh, K. V. (2012). Applying data mining to insurance customer churn management. Int. Proc. Comput. Sci. Inf. Technol, 30, 82-92
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer, 522,523.

- Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons
- Japkowicz, N. (2003, August). Class imbalances: are we focusing on the right issue. In Workshop on Learning from Imbalanced Data Sets II (Vol. 1723, p. 63).
- Visa, S., & Ralescu, A. (2005, April). Issues in mining imbalanced data sets-a review paper In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference (Vol. 2005, pp. 67-73). sn.
- Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
- Staudt, M., Kietz, J. U., & Reimer, U. (1998, August). A Data Mining Support Environment and its Application on Insurance Data. In KDD (pp. 105-111).
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M. (2008). The need to monitor customer loyalty and business risk in the European insurance industry. The Geneva Papers on Risk and Insurance-Issues and Practice, 33(2), 207-218.
- Guillén, M., Nielsen, J. P., Scheike, T. H., & Pérez-Marín, A. M. (2012). Time-varying effects in the analysis of customer loyalty: A case study in insurance. Expert systems with Applications, 39(3), 3551-3558.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of marketing research, 43(2), 204-211