IMPLEMENTACIÓN DE MACHINE LEARNING PARA LA ESTIMACION DEL RIESGO DE FUGA DE LOS CLIENTES EN UNA MARCA DE UNA EMPRESA DE LA INDUSTRIA DEL RETAIL DE MODA EN COLOMBIA

Sebastián Elorza Velásquez

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Diego Fernando Mosquera

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 01 DE 2024

FICHA RESUMEN

TÍTULO: Implementación de machine learning para la estimación del riesgo de fuga de los clientes en empresa de la industria del retail de moda en Colombia

1. ÁREA DE TRABAJO: Dirección de Conocimiento de Clientes en empresa de moda
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): Sebastián Elorza Velásquez
4. CORREO ELECTRÓNICO: sebaselorza@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Calle 38sur # 27-200 Envigado, 3122012501
6. DIRECTOR: Diego Fernando Mosquera
7. VINCULACIÓN DEL DIRECTOR: Cátedra
8. CORREO ELECTRÓNICO DEL DIRECTOR: dfmosquera@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica):
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): GCO
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5): Churn, abandono, machine learning, clientes, retención, recuperación, moda, retail, datos, modelos predictivos, estimación.
13. FECHA DE INICIO: 15/01/2024
14. DURACIÓN ESTIMADA (En meses):12 meses
15. RESUMEN:

RESUMEN

El trabajo de grado presentado, titulado **"Implementación de Machine Learning para la Estimación del Riesgo de Fuga de los Clientes en una Marca de una Empresa de la Industria del Retail de Moda en Colombia"**, tiene como objetivo principal desarrollar una herramienta predictiva que permita identificar los clientes con mayor probabilidad de abandonar la marca. Esto se busca lograr mediante la aplicación de técnicas de machine learning que analicen el comportamiento de los clientes, sus hábitos de compra y las interacciones con la empresa.

El problema central identificado es que la empresa del caso de estudio, Chevignon, sufre una pérdida significativa de clientes cada año, lo que afecta tanto los ingresos como la percepción de marca. En respuesta a esta problemática, se propuso utilizar datos históricos y técnicas de aprendizaje automático para predecir el riesgo de abandono y así mejorar las estrategias de retención.

El modelo de predicción desarrollado emplea varios algoritmos, entre ellos XGBoost, Random Forest, Support Vector Machines (SVM) y redes neuronales artificiales (ANN). Los resultados muestran que el modelo XGBoost obtuvo el mejor desempeño con una precisión del 86.18% y una sensibilidad del 88.35%, lo que lo convierte en la herramienta más adecuada para predecir la fuga de clientes. La capacidad de predecir el abandono permitió a la empresa implementar acciones proactivas, como ofertas personalizadas y programas de fidelización, lo que ayudará a reducir la pérdida de clientes.

El trabajo también enfatiza la importancia de la limpieza y la preparación de los datos, destacando la necesidad de eliminar variables altamente correlacionadas que podrían afectar la precisión del modelo. A lo largo del proceso, se evaluó la importancia de las variables en el modelo, identificándose que factores como la permanencia del cliente y el tiempo en la marca son determinantes en la predicción del abandono.

En cuanto a trabajos futuros, se sugiere continuar optimizando los modelos mediante la incorporación de nuevas variables, el ajuste de hiperparámetros y la experimentación con otros algoritmos, como redes neuronales profundas o técnicas de ensamblado de modelos. También se propone investigar la posibilidad de implementar el modelo en tiempo real y personalizar las estrategias de retención en función del perfil y comportamiento de los clientes.

En conclusión, este trabajo ofrece una herramienta valiosa para la marca, que, al predecir el riesgo de abandono, permitirá a la empresa tomar decisiones más informadas y estratégicas para mejorar la retención de clientes, reduciendo costos asociados y aumentando la competitividad en un mercado en constante cambio.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La pérdida de clientes es un desafío significativo para las empresas de la industria del retail de moda en Colombia. Según un estudio de la consultora Bain & Company, el costo de adquirir un nuevo cliente es entre cinco y siete veces mayor que el costo de retener a uno existente. Por lo tanto, es crucial para las empresas desarrollar estrategias efectivas para la retención de clientes.

El abandono de clientes no solo afecta los ingresos directos de una empresa, sino que también impacta la percepción de la marca y la lealtad del consumidor. Además, amenaza las estrategias a largo plazo y la capacidad de la empresa para mantener una ventaja competitiva en un mercado en constante evolución.

En GCO, compañía que tiene la representación de varias marcas de moda en el país y cuenta con distribución tanto en tiendas físicas como tiendas virtuales, existen áreas de relacionamiento de clientes (CRM) desde hace aproximadamente 6 años. Dichas áreas son las encargadas de generar engagement con los clientes para ofrecerles una mejor experiencia en su paso por la marca, por eso se tiene el objetivo de retener y recuperar el máximo de clientes posibles, brindando experiencias memorables a cada uno de ellos. La marca Chevignon, del grupo GCO, es una empresa de moda con más de 30 años en el mercado colombiano, actualmente está en 38 ciudades del país a través de tiendas propias y franquicias y llega a todo el territorio nacional con su canal online. La marca es referente en el mundo del jeanswear y en prendas de cuero.

En este contexto, el uso de técnicas de Machine Learning ofrece una oportunidad valiosa para predecir el riesgo de abandono de los clientes. La capacidad de identificar a aquellos con mayor probabilidad de abandonar permite a las empresas tomar medidas proactivas para retenerlos mediante estrategias personalizadas, como incentivos o promociones.

El cliente al que se hace referencia en este trabajo es el cliente final, el usuario de las tiendas que compra las prendas de vestir, el consumidor que compra a través de los canales digitales o visita las tiendas físicas y accede a los productos ofrecidos por la marca.

El objetivo general de este proyecto fue desarrollar una herramienta que permita predecir el riesgo de abandono en la industria del retail de moda. Para lograrlo, se recopiló información relevante sobre los clientes, sus hábitos de compra y sus interacciones con la marca. A través del análisis exploratorio de datos, se identificaron los factores relacionados con el riesgo de abandono y se construyeron modelos de Machine Learning para predecir dicho riesgo.

Los resultados de este proyecto incluyen una comprensión más profunda de los clientes y un modelo de aprendizaje automático que permite tener una mayor certeza en las características que puede tener el potencial cliente a abandonar y mayores insumos del posible riesgo de abandono de los clientes, esto ayuda a enfocar las áreas de mercadeo en el desarrollo de estrategias de retención más efectivas. Esto no solo ayudó a reducir la pérdida de clientes, sino que también mejoró la rentabilidad y la sostenibilidad de las empresas en el competitivo mercado de la moda.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

GCO, específicamente su marca Chevignon es un productor y distribuidor de diferentes prendas y productos de moda en Colombia y varios países de América Latina, donde ha visto un golpe constante debido al gran número de clientes que compran en sus tiendas pero que tiempo después no vuelven a comprar (abandono), es por esto que toma importancia apalancarse en la ciencia de datos para resolver un problema constante y que golpea fuertemente las ventas y las finanzas de la compañía.

La pérdida masiva de clientes no solo impacta los ingresos directos, sino que conlleva una pérdida en el valor de la marca y la lealtad del consumidor. Además, este fenómeno pone en riesgo las estrategias a largo plazo de las compañías y su capacidad para mantener una ventaja competitiva sostenible en un mercado de moda en constante evolución.

Conscientes de la importancia crucial de retener a los clientes existentes, la marca implementó estrategias innovadoras y proactivas para abordar este desafío.

El desafío que se abordó en este estudio se centra en la posible pérdida de clientes que enfrenta la empresa. Se propuso explorar la problemática asociada a la deserción de clientes. Se identificó y comprendió a fondo los factores que contribuyen a que los consumidores abandonen la empresa, y así, se crearon alertas tempranas que permiten a la compañía abordar de manera proactiva y personalizada la estrategia de retención de clientes.

1.2. FORMULACIÓN DEL PROBLEMA

El planteamiento del problema implica responder a los siguientes cuestionamientos: ¿Es posible estimar con machine learning el riesgo que tiene un cliente de abandonar una marca de moda?, ¿Cuáles son los principales factores que contribuyen al abandono de los clientes? ,¿Cuáles modelos de ciencia de datos permiten ayudar a entender (predecir) este comportamiento?, Y finalmente, ¿Cómo evaluar el desempeño de los modelos de predicción que nos ayude con la estimación de la probabilidad de riesgo de abandono en una industria como el retail moda?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar una herramienta que permita estimar el riesgo de abandono de un cliente en la marca Chevingon con uso de ciencia de datos.

2.2 OBJETIVOS ESPECÍFICOS

- Desarrollar una herramienta con machine learning para predecir la probabilidad de abandono de un cliente en una marca de moda.
- Seleccionar las técnicas de Machine Learning que se puedan usar para resolver este tipo de necesidades.
- Entrenar modelos de machine learning con los datos obtenidos que permitan predecir el riesgo de abandono.
- Evaluar el performance de los modelos creados.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 ANTECEDENTES

Trabajos relacionados sobre el abandono de clientes:

- **Modelo de predicción de abandono en la banca minorista utilizando Algoritmo C-Means:** El artículo presenta un modelo basado en métodos difusos para la predicción del abandono en la banca minorista. Se aplicó un análisis discriminante canónico para revelar variables que proporcionan la máxima separación entre grupos de abandonados y no abandonados. Debido a la naturaleza difusa de los problemas prácticos de gestión de relaciones con los clientes, se demostró que los métodos difusos funcionaban mejor que los clásicos [1]. Este trabajo se relaciona con el trabajo de grado debido a que abre un poco la visión a revisar otro tipo de modelos diferentes a los usados regularmente (Random Forest, Regresión Logística, SVM), ya que da un entendimiento de como los métodos difusos pueden permitir mejores resultados y el análisis discriminante permite entender una separación entre los grupos de abandono y no abandono.

- **Predicción de pérdida de clientes para empresas minoristas:** La pérdida de clientes ocurre cuando un cliente interrumpe su interacción con la empresa. En el negocio minorista, se considera que un cliente es desechado una vez que sus transacciones caducan durante un periodo de tiempo determinado. Este trabajo se enfoca en una tienda minorista de regalos en Reino Unido con clientes principalmente mayoristas, genera agregaciones de los datos para generar conjuntos basados en facturas y clientes, este valor de abandono se determina en función de las transacciones de los clientes y ejecutan 3 algoritmos: Random Forest, SVM, AB boosting [2]. Este trabajo se relaciona con el nuestro porque está en el mismo sector (retail, aunque no es moda) y desarrolla algunos algoritmos que he investigado y pueden ser útiles al momento de clasificar el abandono de nuestros clientes en la industria de la moda.

- **Comparación de algoritmos de aprendizaje profundo para predecir la rotación de clientes dentro de una industria minorista:** Este documento demuestra como a través de los datos transaccionales se crean características que pueden identificarse como importantes para predecir la deserción dentro de la industria minorista. Los datos proporcionados en el trabajo son de un supermercado, por lo tanto, los abandonos identificados y los resultados obtenidos se basan en escenarios reales. Aplican algoritmos de aprendizaje profundo como redes neuronales convolucionales (CNN) y Restricted Boltzmann Machine [3]. Al igual que el trabajo anterior, este nos permite guiarnos para el desarrollo de nuestro proyecto aplicado, teniendo en cuenta que también se hace en el retail y enfocado directamente en el cliente final (a diferencia del anterior que eran

empresas B2B), adicional, nos presenta la implementación en dos algoritmos que aún no se han tenido en cuenta pero que pueden tener un desempeño superior, como lo son las redes neuronales convolucionales (CNN) y el Restricted Boltzmann Machine.

- **Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones:** Este trabajo de grado demuestra el problema tan costoso que es el abandono para una empresa que está en la industria Telco y desarrollaron un modelo predictivo de Churn para identificar proactivamente a los clientes en riesgo de abandonar y tomar medidas para retenerlos. Este trabajo proporciona una guía para la creación de un modelo predictivo de Churn en empresas de telecomunicaciones y demuestra el potencial de la tecnología Machine Learning para mejorar la retención de clientes. [4]
- **Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático:** La industria de las telecomunicaciones está adoptando cada vez más sistemas digitales de gestión de relaciones con clientes, y la predicción de la pérdida de clientes es crucial. Este estudio propone un modelo mejorado para predecir la rotación de clientes utilizando técnicas de minería de datos. Se analizaron datos históricos para identificar patrones que indiquen clientes en riesgo de abandono. Se utilizaron algoritmos como el análisis de regresión, árboles de decisión y redes neuronales artificiales. Los resultados demostraron que el modelo propuesto, con la incorporación del impulso, superó a los modelos tradicionales en la predicción precisa de la rotación de clientes. Este estudio destaca la importancia de la minería de datos y el aprendizaje automático para mejorar la retención de clientes en la industria de las telecomunicaciones. [5]
- **Modelo de Churn para retención de clientes de Seguros Voluntarios:** La tasa de cancelación de clientes en seguros voluntarios es un problema significativo para las compañías financieras en Colombia. Adquirir nuevos clientes es más costoso que retener los existentes, por lo que es crucial predecir y prevenir la cancelación. Este estudio propone un modelo para predecir la cancelación de clientes utilizando aprendizaje automático y minería de datos. Se analizan datos históricos para identificar patrones que indiquen clientes en riesgo de cancelación. Se utilizan diferentes algoritmos de aprendizaje automático, como la regresión logística, los árboles de decisión y las redes neuronales artificiales. Los resultados demuestran que el modelo propuesto puede predecir con precisión la cancelación de clientes. Este estudio destaca la importancia del aprendizaje automático y la minería de datos para mejorar la retención de clientes en la industria financiera. [6]

3.2 MARCO TEÓRICO

3.2.1 Contexto organizacional

Se realiza una búsqueda de referentes al desarrollo de modelos para predecir el abandono en diferentes industrias, adicional se aborda un poco la problemática que está teniendo la organización actualmente con el problema de deserción de clientes y se presentan a continuación algunos temas relacionados:

En GCO, Compañía que tiene la representación de varias marcas de moda en el país y cuenta con distribución tanto en tiendas físicas como tiendas virtuales, cuenta con áreas de relacionamiento de clientes (CRM) desde hace aproximadamente 6 años. Dichas áreas son las encargadas de generar engagement con los clientes para ofrecerles una mejor experiencia en su paso por la marca. Este proyecto se enfoca en una de las marcas principales: Chevignon.

En el dinámico y competitivo mercado de la moda en Colombia, la marca enfrenta un desafío crítico: la constante pérdida de clientes. A pesar de haber logrado establecer una base considerable de casi 750 mil clientes, la empresa sufre la desafortunada situación de perder casi el 50% de ellos anualmente.

Esta disminución significativa en la base de clientes representa una amenaza seria para la estabilidad y el crecimiento de la compañía. La constante salida de clientes no solo impacta negativamente en los ingresos, sino que también compromete la posición competitiva en el mercado. La fidelización de los consumidores se vuelve crucial, ya que la adquisición de nuevos clientes resulta ser un proceso sumamente costoso, es aproximadamente cinco veces más costoso que retener a los clientes actuales.

El abandono o churn en la marca se define como la persona que deja de comprar después de 360 días desde su última compra y la tasa de abandono es la cantidad de clientes que se van perdiendo cada mes sobre el total de activos que tiene la compañía. La industria retail moda se caracteriza por una alta tasa de rotación de clientes (churn o abandono) [7]. Esto se debe a una serie de factores, como la competencia, la facilidad de acceso a la información y la variedad de opciones disponibles para los consumidores. El abandono es un problema importante para las empresas retail moda, ya que puede tener un impacto negativo en los ingresos, la rentabilidad y la satisfacción del cliente. Por ello, es importante desarrollar estrategias para reducir la tasa de abandono.

Una de las estrategias más efectivas para reducir el abandono es la predicción de abandono. La predicción de abandono permite a las empresas identificar a los clientes que tienen mayor probabilidad de dejar de ser clientes. Una vez identificados estos clientes, las empresas pueden implementar acciones para retenerlos.

La predicción de abandono puede realizarse utilizando diferentes métodos, como el análisis de datos, el machine learning y la inteligencia artificial.

3.2.2 Análisis de datos

El análisis de datos es un proceso sistemático que implica la recopilación, organización, interpretación y presentación de datos con el objetivo de descubrir patrones, tendencias y relaciones significativas. A través de diversas técnicas estadísticas y de aprendizaje automático, se extrae información valiosa de grandes volúmenes de datos, lo que permite tomar decisiones más informadas y basadas en evidencias. El análisis de datos se ha convertido en una herramienta fundamental en diversos campos, desde las ciencias sociales y la medicina hasta los negocios y la ingeniería, impulsando la innovación y el desarrollo en la era de la información [8].

3.2.3 Inteligencia artificial

La inteligencia artificial es un campo de la informática que se ocupa del desarrollo de sistemas capaces de pensar y actuar como los humanos. La inteligencia artificial se puede utilizar para desarrollar modelos de predicción que sean aún más precisos que los modelos basados en machine learning [9].

Algunos de los sistemas de inteligencia artificial más utilizados para la predicción de abandono son los sistemas de aprendizaje automático reforzado y los sistemas de procesamiento de lenguaje natural.

3.2.4 Machine learning

- Es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos que permiten a las máquinas aprender y mejorar a partir de datos sin ser explícitamente programadas para ello. A través de diferentes técnicas como el aprendizaje supervisado, no supervisado y por refuerzo, las máquinas pueden identificar patrones, hacer predicciones y tomar decisiones basadas en la información que procesan [10]. El machine learning se puede utilizar para desarrollar modelos de predicción que sean más precisos que los modelos basados en análisis de datos.
- Algunos de los algoritmos de machine learning más utilizados para la predicción de abandono son los árboles de decisión, los modelos de regresión logística y los modelos de aprendizaje profundo.

El aprendizaje supervisado es un subcampo del aprendizaje automático que se ocupa de la construcción de modelos que aprenden a partir de datos etiquetados. En el aprendizaje supervisado, se proporciona al modelo un conjunto de datos de entrenamiento que contiene ejemplos de la variable de entrada y la variable de salida. El modelo aprende a relacionar las variables de entrada con la variable de salida a partir de este conjunto de datos.

Algunas técnicas de aprendizaje supervisado

Existen muchas técnicas de aprendizaje supervisado, cada una con sus propias ventajas y desventajas. Algunas de las técnicas de aprendizaje supervisado más utilizadas son las siguientes:

- **Máquinas de vectores de soporte (SVM):**Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) son un tipo de algoritmo de aprendizaje supervisado utilizado principalmente para tareas de clasificación y regresión. SVM fue introducido por Vladimir Vapnik y sus colegas en la década de 1990, y se ha convertido en una herramienta poderosa en el campo del aprendizaje automático debido a su eficacia en la clasificación de datos complejos [11].

  El objetivo principal de una SVM es encontrar un hiperplano en un espacio multidimensional que separe las distintas clases de datos con el margen más amplio posible. Este hiperplano es conocido como el hiperplano de separación óptimo. Los vectores de soporte son los puntos de datos más cercanos a este hiperplano, y son los que definen la posición y orientación del mismo.

  Para comprender cómo funciona una SVM, es útil considerar un caso de clasificación binaria (donde hay dos clases). En este contexto, el objetivo de la SVM es encontrar un hiperplano (una línea en dos dimensiones, un plano en tres dimensiones, etc.) que separe los puntos de datos de una clase de los puntos de datos de la otra clase, maximizando al mismo tiempo la distancia (margen) entre los puntos de datos más cercanos de cada clase y el hiperplano [12].

- **Regresión logística:** Es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación. La regresión logística predice la probabilidad de que una observación pertenezca a una clase determinada [8]. Es especialmente útil cuando se trabaja con variables dependientes binarias (es decir, aquellas que tienen dos posibles resultados, como "sí/no", "verdadero/falso", o "1/0") [13]. Este método es ampliamente utilizado en situaciones donde se necesita clasificar observaciones en dos categorías.
- **Random forest:** Es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación y regresión. El algoritmo de Random Forest construye múltiples árboles de decisión durante el entrenamiento y los utiliza para producir una predicción agregada. Cada árbol en el bosque es entrenado en una muestra diferente del conjunto de datos original mediante una técnica conocida como *bagging* (bootstrap aggregating) [14]. En este proceso, se selecciona aleatoriamente un subconjunto de características para construir cada árbol, lo que reduce la correlación entre los árboles individuales y mejora la capacidad generalizadora del modelo.

  Una vez que el bosque de árboles de decisión ha sido construido, la predicción final para un nuevo dato se realiza mediante un proceso de votación en el caso de clasificación, o mediante la media de las predicciones individuales de los árboles en el caso de regresión [15]. Este enfoque de ensamble mitiga el problema del sobreajuste que es común en los árboles de decisión individuales, donde el modelo tiende a aprender en exceso los detalles específicos del conjunto de datos de entrenamiento.

- **Gradient boosting:** Es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación y regresión. Gradient boosting construye un conjunto de árboles de decisión secuencialmente, cada uno de los cuales intenta corregir los errores del árbol anterior [16].

  El principio central de *Gradient Boosting* es construir un modelo fuerte a partir de una serie de modelos débiles, típicamente árboles de decisión. Cada modelo sucesivo intenta corregir los errores cometidos por el modelo anterior, entrenándose sobre los residuos, es decir, las diferencias entre las predicciones del modelo actual y los valores reales [16]. Esto se logra minimizando una función de pérdida específica mediante el descenso del gradiente, de ahí el nombre "Gradient Boosting".

  El proceso comienza con la creación de un modelo inicial simple, y luego se construyen iterativamente nuevos modelos que se agregan para corregir los errores residuales de los modelos anteriores [17]. En cada iteración, los nuevos modelos se entrenan para predecir los residuos del modelo anterior, y la combinación de todos estos modelos es lo que forma la predicción final.

- **Redes neuronales artificiales:** Las redes neuronales artificiales (ANN, por sus siglas en inglés) son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano, compuestas por unidades denominadas neuronas artificiales que se organizan en capas. Estas redes son capaces de aprender a partir de datos mediante un proceso de ajuste de pesos en las conexiones entre neuronas. Este aprendizaje se realiza a través de algoritmos como la retropropagación, permitiendo a las redes reconocer patrones complejos en los datos y realizar tareas como clasificación, regresión y predicción [18]. Las ANN son ampliamente utilizadas en problemas de reconocimiento de voz, visión por computadora, procesamiento de lenguaje natural y otros campos debido a su capacidad para modelar relaciones no lineales en grandes volúmenes de datos [19].

3.2.5 Evaluación del desempeño de los modelos

El desempeño de un modelo de aprendizaje supervisado se puede evaluar utilizando una variedad de métricas, como la precisión, la sensibilidad, la especificidad y la curva ROC. Es importante partir como base inicial el cálculo de la matriz de confusión.

Matriz de Confusión: Es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para describir el rendimiento del modelo de clasificación (o "clasificador") sobre un conjunto de datos de prueba cuyos valores reales se conocen.

En la figura 1 se observa cómo se distribuyen las clasificaciones de los modelos para crear la matriz de confusión.

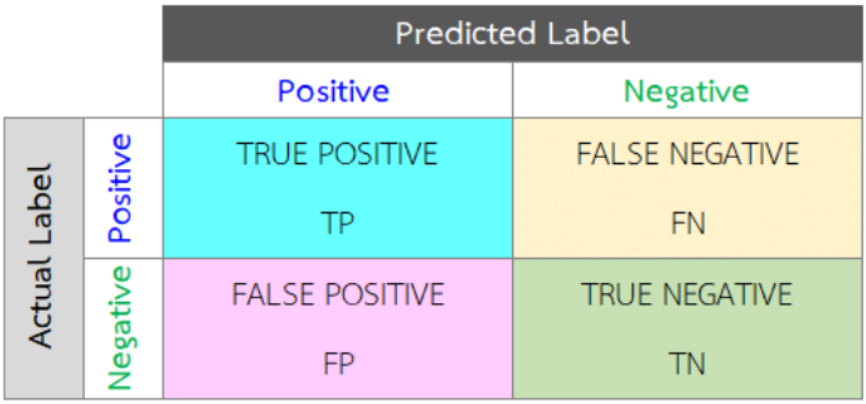

Figura 1Matriz de Confusión

Cada predicción puede ser uno de cuatro resultados, basado en cómo coincide con el valor real:

- **Verdadero Positivo (TP):** Predicho Verdadero y Verdadero en realidad**.**
- **Verdadero Negativo (TN):** Predicho Falso y Falso en realidad.
- **Falso Positivo (FP):** Predicción de verdadero y falso en la realidad.
- **Falso Negativo (FN):** Predicción de falso y verdadero en la realidad.

Ahora si describimos cada métrica y cómo se calcula partiendo de los resultados de la matriz de confusión.

Precisión (Accuracy): La precisión es la proporción de observaciones correctamente clasificadas [20].

$$Accuracy = TP \ / \ Predicción\ Si$$

Sensibilidad (Recall): La sensibilidad es la proporción de observaciones de la clase positiva que se clasifican correctamente [20].

$$Recall = TP \ / \ (TP + FN)$$

Especificidad: La especificidad es la proporción de observaciones de la clase negativa que se clasifican correctamente [20].

$$Especificidad = VN \ / \ \ Predicciones\ No$$

Curva ROC: La curva ROC es una gráfica que representa la sensibilidad en función de la especificidad para diferentes umbrales de clasificación [20].

La elección de la métrica de evaluación adecuada depende del problema específico que se está abordando. Por ejemplo, si el objetivo es minimizar el número de falsos negativos, se puede utilizar la sensibilidad como métrica de evaluación.

En el caso de la predicción de abandono, una métrica de evaluación adecuada podría ser la tasa de falsos negativos. Esto se debe a que es importante identificar a los clientes que tienen mayor

probabilidad de abandonar, incluso si esto significa que se etiquetan incorrectamente a algunos clientes como "clientes que abandonarán" que en realidad no lo harán.

4. METODOLOGÍA, ESTRUCTURACIÓN DE LA INFORMACIÓN, PROCESAMIENTO DE DATOS Y MODELACIÓN

Para la ejecución del proyecto se siguió la metodología CRISP DM.

CRISP DM es un acrónimo de Cross Industry Standard Process for Data Mining, que en español significa Proceso estándar de la industria para la minería de datos. Es una metodología de desarrollo de proyectos de minería de datos que se ha convertido en el estándar de facto en la industria.
CRISP DM se divide en seis fases:

1. **Comprensión del negocio:** En esta fase, se recopila información sobre el negocio y el problema que se desea resolver con la minería de datos. Se identifican los objetivos del proyecto y los requisitos del cliente [21].
2. **Entendimiento de datos**: En esta fase, se exploran los datos para identificar patrones y tendencias. Se utilizan técnicas de análisis descriptivo para visualizar los datos y comprender su estructura [23].
3. **Preparación de los datos:** En esta fase, se preparan los datos para su análisis. Se realizan tareas como la limpieza de datos, la integración de datos y la transformación de datos [22].
4. **Modelado:** En esta fase, se desarrollan modelos para predecir o explicar los datos. Se utilizan técnicas de aprendizaje automático para construir modelos que sean capaces de generalizar a nuevos datos [8].
5. **Evaluación:** En esta fase, se evalúan los modelos desarrollados. Se utilizan métricas de evaluación para determinar la precisión y la efectividad de los modelos [24].
6. **Implementación:** En esta fase, se implementan los modelos desarrollados en un entorno productivo. Se realizan tareas como la integración de los modelos en los sistemas existentes y la formación del personal para su uso [25].

Ventajas de CRISP DM

- Es una metodología flexible que puede adaptarse a diferentes tipos de proyectos de minería de datos.
- Proporciona un marco estructurado para el desarrollo de proyectos exitosos.
- Ayuda a los profesionales de la minería de datos a evitar errores comunes.

Desventajas de CRISP DM

- Puede ser compleja de implementar en proyectos grandes o complejos.

- Requiere la participación de un equipo multidisciplinario de profesionales.

En la figura 2 se muestra el flujo y las fases de los procesos para la metodología CRISP-DM que se utilizará.

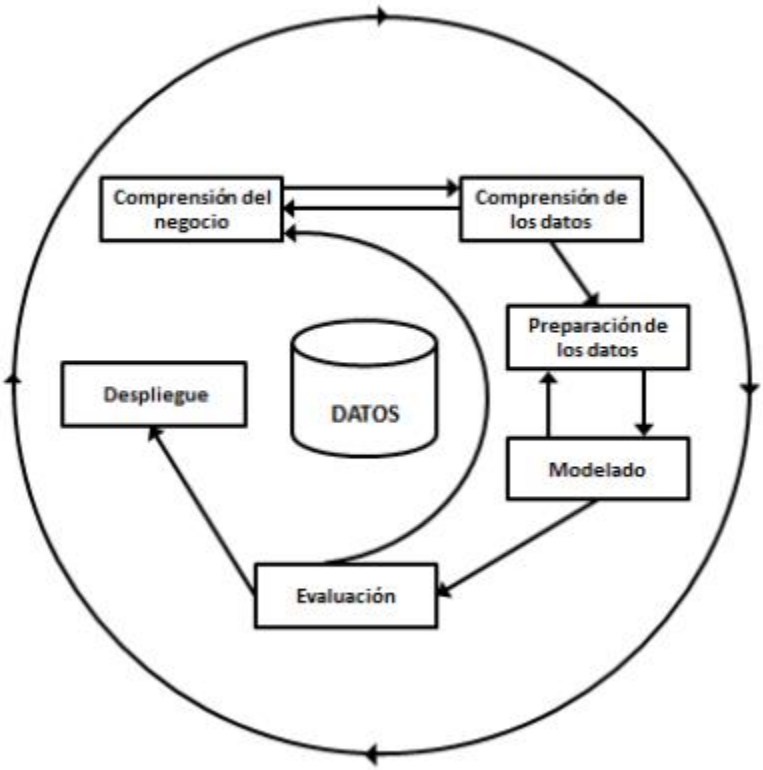

Figura 2 Fases del modelo de proceso de la metodología CRISP-DM

4.1 Comprensión del negocio

La pérdida de clientes en el sector del retail moda es un problema que constantemente enfrentan las compañías que hacen parte de esta industria.

La pérdida masiva de clientes no solo impacta los ingresos directos, sino que también conlleva una pérdida en el valor de marca y lealtad del consumidor. Además, este fenómeno pone en riesgo las estrategias a largo plazo de las compañías y su capacidad para mantener una ventaja competitiva sostenible en un mercado de moda en constante evolución.
GCO, específicamente su marca Chevignon, es una empresa de moda con más de 30 años en el mercado colombiano, actualmente está en 38 ciudades del país a través de tiendas propias y franquicias y llega a todo el territorio nacional con su canal online. La marca es referente en el mundo del jeanswear y en prendas de cuero.

La marca cuenta con clientes activos (compradores en el mismo año) de 750.000 aproximadamente, y al año ve como casi 400.000 clientes (equivalente a cerca del 53%) dejan de comprar en ella en un solo año, eso podría tener un impacto en ventas de miles de millones.

Actualmente la organización cuenta con información demográfica y transaccional anonimizada de los clientes que podría permitir reconocer algunos patrones en su comportamiento de compra. La marca considera un cliente “dormido o inactivo” si deja de comprar en 12 meses (recencia mayor o igual a 365 días), lo que hace que los esfuerzos de retención se centren en clientes con recencias cercanas a este dato, pero cada cliente puede tener motivaciones y formas diferentes de comprar, por eso la idea fue construir un modelo que permitió entender muchos más atributos o puntos de datos y no solo sea la recencia la que genere la “alerta” de abandono de los clientes.

El grupo decidió crear las áreas de CRM y la gerencia de clientes en 2016 aproximadamente para tener un área responsable de la correcta gestión de los clientes, entender sus necesidades y apoyarse en los datos para tomar decisiones. En esa búsqueda de entregar una mejor experiencia a los clientes se empezaron a realizar estrategias de mercadeo para retener el cliente, atraer nuevos compradores y los que ya hacen parte de la organización aumenten sus frecuencias de compras, así que se creó una métrica que se llamó “Cliente Activo” y se definió como regla de negocio que serían las personas que tenían al menos una compra en los últimos 12 meses. Con esa regla de negocio establecida, se empezó a evidenciar que año tras año un gran número de personas dejaba de comprar o se “inactivaba”, la tasa promedio de clientes inactivos en el último año por ejemplo fue del 53%, siendo casi igual a la cantidad de clientes nuevos que ingresan a la compañía. Viendo el panorama en cifras y con el conocimiento que se tiene en inversión sobre un cliente nuevo, se fue creando la necesidad de “atacar” a los clientes que se iban pero siempre se ha hecho de una forma muy reactiva, solo después de que este abandona la compañía se activan estrategias para tratar de recuperarlos, ya que mientras están en ella, las áreas de relacionamiento con el cliente generan estrategias de mercadeo relativamente trasversales y no tan enfocadas en una persona con riesgo de fuga. De acuerdo a la tasa promedio histórica de cantidad de inactivos que se tiene cada año (50% en promedio) y representando una cifra muy significativa en ventas, se observa la necesidad de analizar más a detalle esta problemática, tratar de entender qué patrones tenían los clientes que estaban dejando de comprar, de ahí surge la necesidad de aprovechar los datos que se tienen para intentar adelantarse al abandono del cliente, inicialmente sin mucho éxito, ya que se estaba usando de manera poco estructurada y sin modelos estadísticos o metodológicas en ciencia de datos que pudieran apalancar este proceso. Desde el 2022 se viene trabajando en la estructuración de un corpus de datos que permita poder anticiparse a la fuga de los clientes, este es el primer paso y el punto de partida para el objetivo de poder crear un modelo de machine learning que permita predecir el riesgo de abandono de los clientes y así poder ser más asertivos en las campañas y estrategias de recuperación que se plantean desde las áreas de mercadeo.

Las gráficas a continuación proporcionan un mayor entendimiento de la información demográfica de los compradores.

En la figura 3 encontramos que la mayoría de clientes son hombres (47%) y que hay un número significativo de personas que no tienen género registrado.

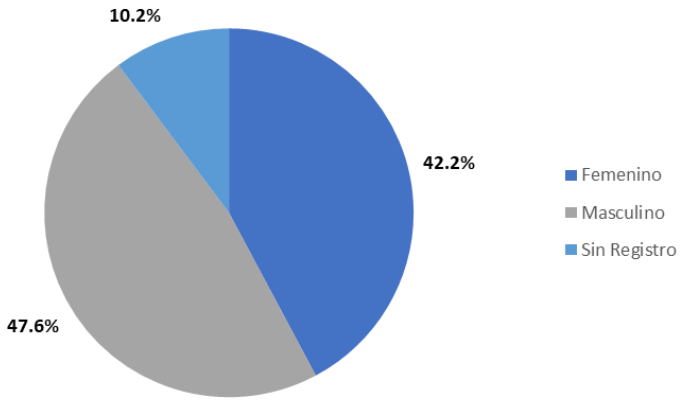

Figura 3 Distribución de clientes por género

En la figura 4 vemos que casi la mitad de las compras (48.9%) se realiza a través de las tiendas propias y el 20% se hace por el canal online.

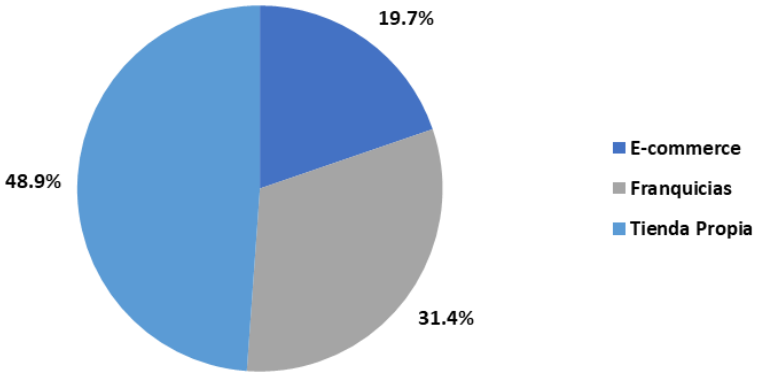

Figura 4 Distribución de clientes por canal de venta

En la figura 5 se observa que no hay una edad que sea predominantemente fuerte, la mayoría de los clientes está entre los 30 y los 50 años (64%).

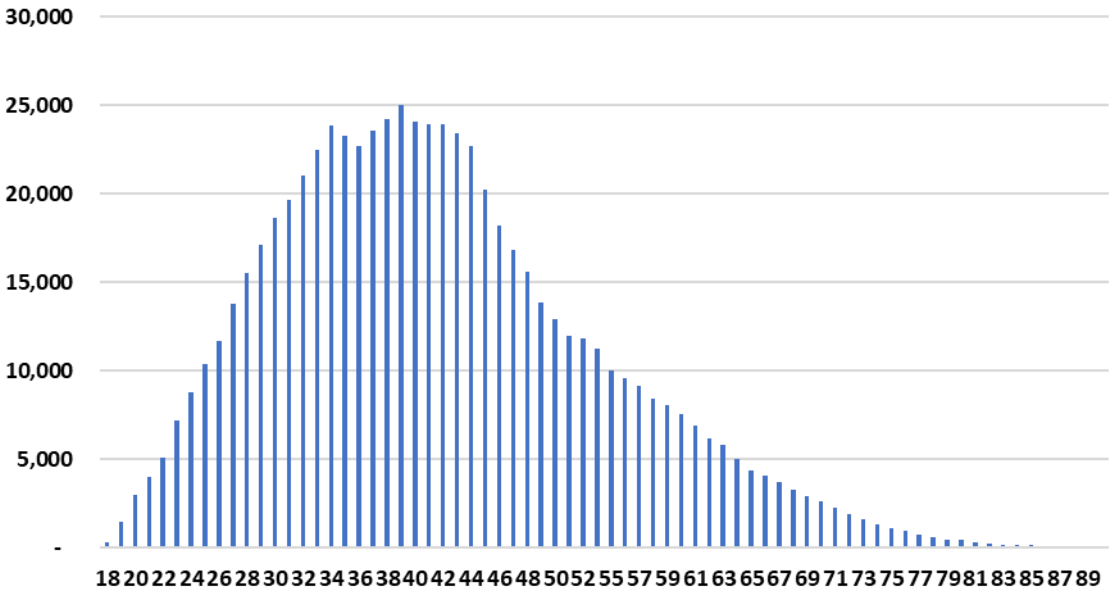

Figura 5 Distribución de la edad de los clientes

En la figura 6 se observa que las ciudades que más clientes tienen son Medellín y Bogotá con el 40% entre ambas. Anexamos gráfico del top 10 de ciudades.

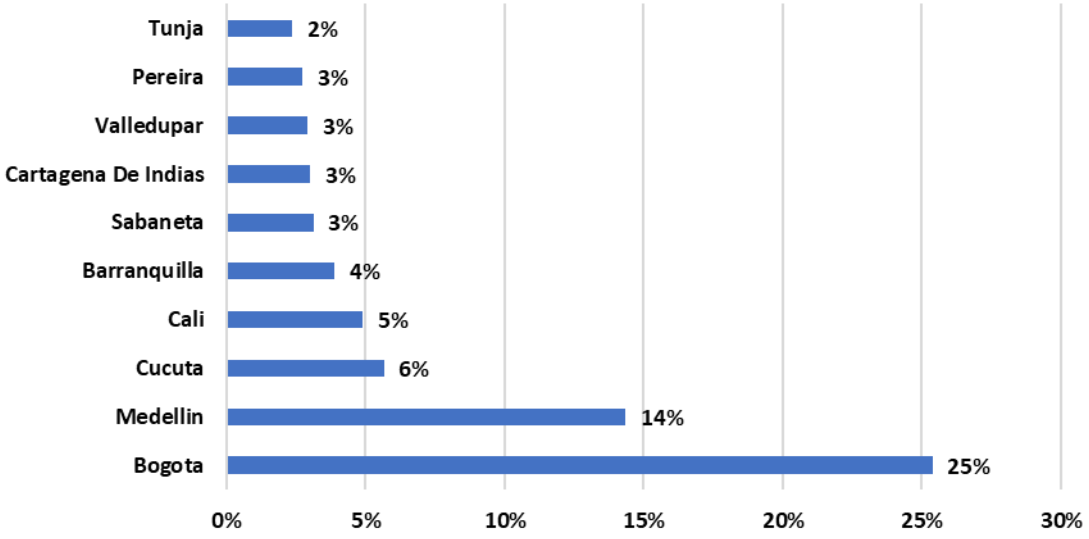

Figura 6 Top 10 Ciudades con mayor número de clientes

A continuación, vemos una gráfica donde muestra la evolución de los inactivos y la relación con la cantidad de clientes activos año tras año desde 2016. Se observa que el comportamiento ha venido creciendo en clientes inactivos después de la pandemia (2021 en adelante) y la tasa de nuevos o recuperados decrece justamente también post pandemia.

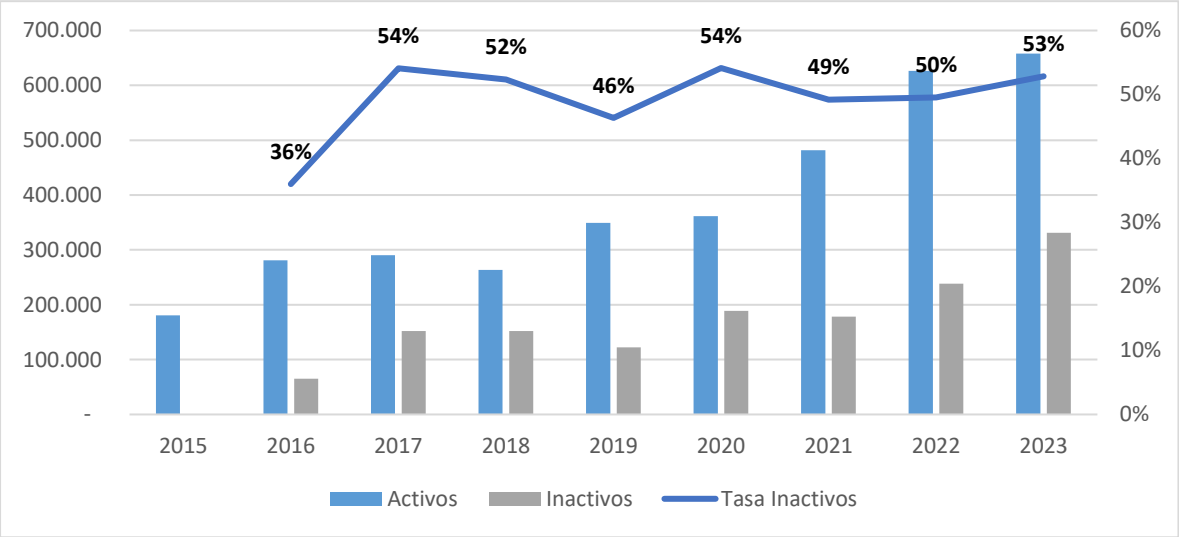

Figura 7 Evolución de clientes y su estado

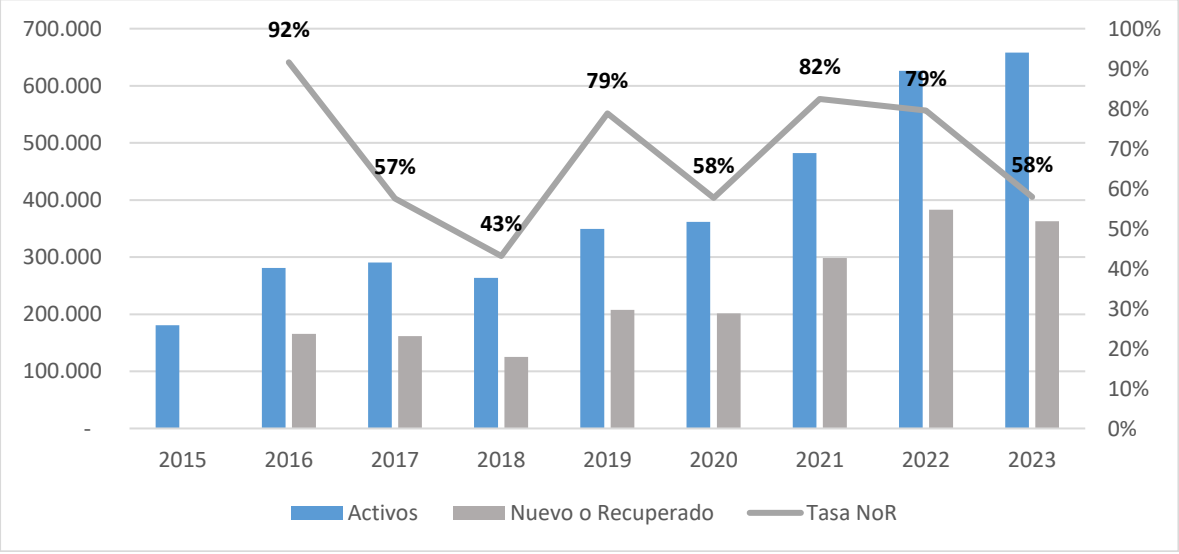

Figura 8 Evolución de clientes nuevos y recuperados

De acuerdo a la figura 8, se observa que la tasa de clientes nuevos y recuperados viene cayendo desde hace varios años. En resumen, en la marca se tiene un abandono anual de aproximadamente 400.000 personas, correspondiente al 58% de los compradores del año.

4.1.1 Recolección de los Datos

Dentro de la organización se tienen diferentes tipos de datos y atributos del cliente que se espera puedan ser útiles al momento de estimar el riesgo de fuga de un cliente de una marca. La información está almacenada dentro de los servidores on-premise de la compañía, se recolectan a nivel del POS de facturación en tienda física, se tiene también información de las transacciones online que realiza, adicional, por ser parte del proceso de marketing relacional, también se tiene información demográfica que puede ser de utilidad para estos casos. Todos estos diversos atributos se almacenan en una base de datos y se consulta a través de SQL o de un Cubo OLAP construido para tal fin. El uso y la recolección de los datos de los clientes se hace bajo la Ley 1581 de 2012 (Ley de protección de datos personales).

4.2 Entendimiento de los Datos

Presentamos un diccionario con los datos que tenemos actualmente y que fueron usados para crear el corpus y el modelo de predicción.

Tabla 1 Diccionario de variables

Nombre del campo	Tipo de Dato	Explicación
ID cliente	String	Id único del cliente (Diferente a la cédula)
Genero Cliente	String	Género del Cliente (Femenino, Masculino, NA)
Edad	String	Edad del cliente
Canal (TP)	String	Canal Preferido de compra del cliente (Tienda Física u Online)
Ciudad Tienda (TP)	String	Ciudad de la tienda preferida
Fecha (Reg)	Date	Fecha de registro del cliente en la marca
Vlr Venta Sin Iva	Numérico	Valor de la venta sin iva
Cantidad Facturas	Numérico	Cantidad de facturas en el periodo definido
Cantidad Devoluciones	Numérico	Cantidad de devoluciones realizadas por el cliente en el periodo definido
Frecuencia	Numérico	Cuántas veces visita el cliente la tienda y realiza una compra
Unidades Vendidas	Numérico	Total unidades vendidas en el periodo
Unidades Devueltas	Numérico	Total unidades devueltas en el periodo
Recencia	Numérico	Días desde la última compra del cliente
Cupo_Tarjeta	Numérico	Valor del cupo aprobado de a tarjeta
Estado_Tarjeta	String	Estado actual de su tarjeta (Cancelado, Activo, etc)
Cupo_Cliente	Numérico	Valor del crédito SU+Pay
Saldo_Puntos	Numérico	Saldo Puntos SU+ (Programa de fidelización de la compañía)
CLTV	Numérico	Ciclo de Vida del Cliente, se calcula a 3 años
frecuencia_mediopago	Numérico	Cantidad de veces que usa un medio de pago
Permanencia	Numérico	Es un campo calculado donde nos indica si el cliente ha comprado
tiempo_marca	Numérico	Días desde que el cliente se registró en la marca
Mes_UV	String	Mes de Ultima Venta
Churn	Booleano	Si el cliente está activo o inactivo en la marca

En total se tienen 23 variables, con información transaccional recolectada desde enero del 2018 y consolidada con la cantidad de clientes que tienen al menos una compra desde el periodo definido.

La información que se tiene actualmente para construir el modelo de abandono se construyó a través de una serie de consultas por medio de SQL, ya que el corpus de datos está almacenado en los servidores *on-premise* de la compañía. Compartimos un diagrama inicial de cómo están distribuidos los datos en tablas de hechos y dimensiones separadas que al final se unen a través de consultas SQL o Cubos OLAP.

En la figura 9 se observa el modelo semántico del cubo que contiene los datos que se utilizaron en el modelo.

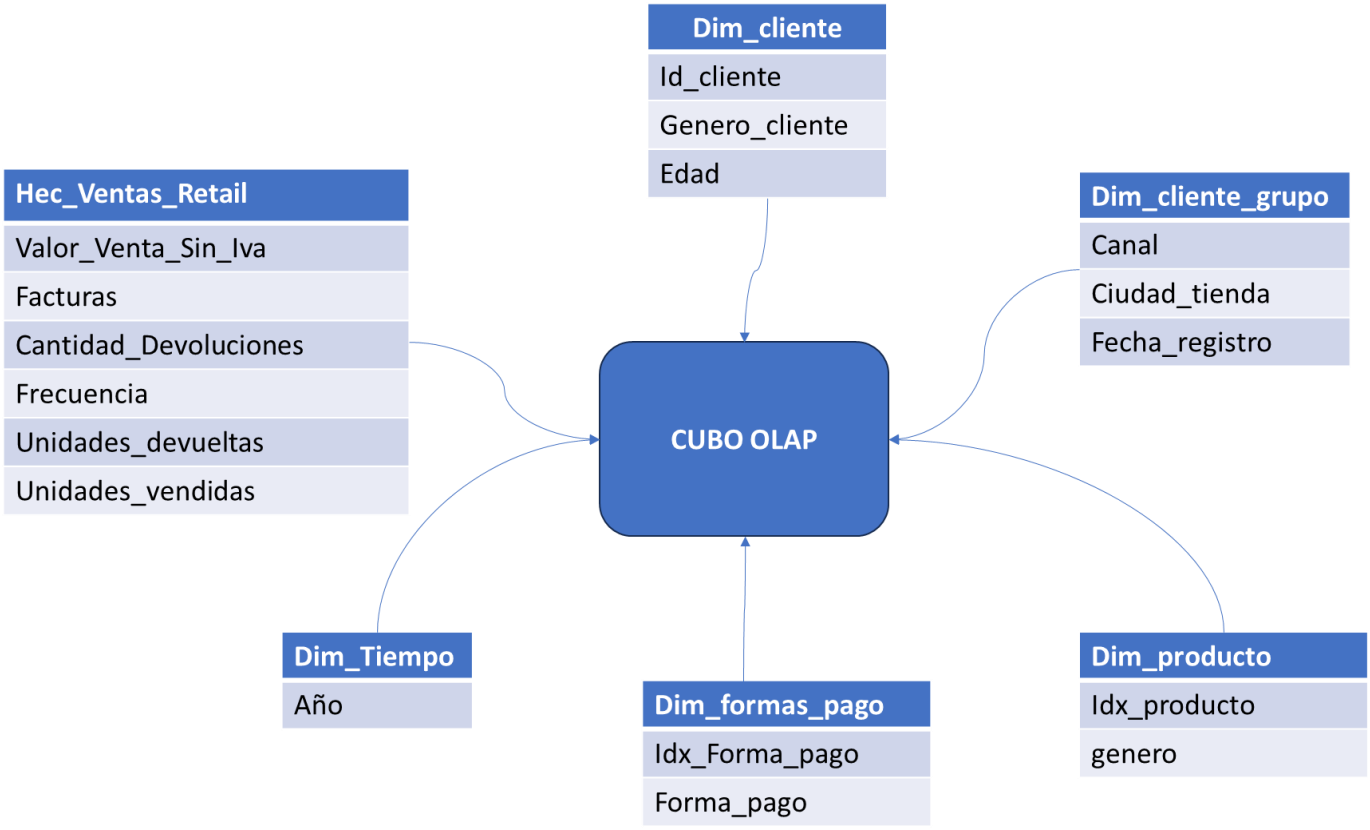

Figura 9 Estructura OLAP

Cabe resaltar que en el 2023 se tuvo un gran número de PQRs (Cerca de 87 mil) con todo tipo de tipificaciones y muchas de ellas no tienen como asociarse a los clientes, ya que no tienen identificación para hacerlo, lo que dificulta relacionarlo a las personas que abandonan la marca.

4.2.1 Análisis Exploratorio de Datos (EDA)

Se realizó un análisis exploratorio de datos en Python con la ayuda de la librería sweetviz para tener un mayor entendimiento de la información del dataset, cómo está distribuida, que datos faltantes tenemos, cómo se correlacionan las variables, entre otros.

Tenemos un dataset con 759.800 filas y 23 variables, de las cuales, 6 son categóricas, 15 numéricas y 2 de texto.

- **ID_cliente**: Es un código único que se crea de cada cliente cuando registra una venta por primera vez, permite identificarlo a lo largo de su estancia en la organización.

- **Género:** Es el género del cliente, tenemos Masculino, Femenino y NA que se aplica cuando no se tiene como poblar el campo, la mayoría de los clientes sin género son compradores online, ya que la plataforma actual no solicita dicho campo. En el dataset a utilizar, del 100% de los clientes con género prima los datos masculinos con un 53% seguidos por los femeninos con un 47%, también es importante tener en cuenta que 10% de los clientes no registran género. En la figura 10 se muestra la distribución de la cantidad de clientes (porcentualmente) por el género.

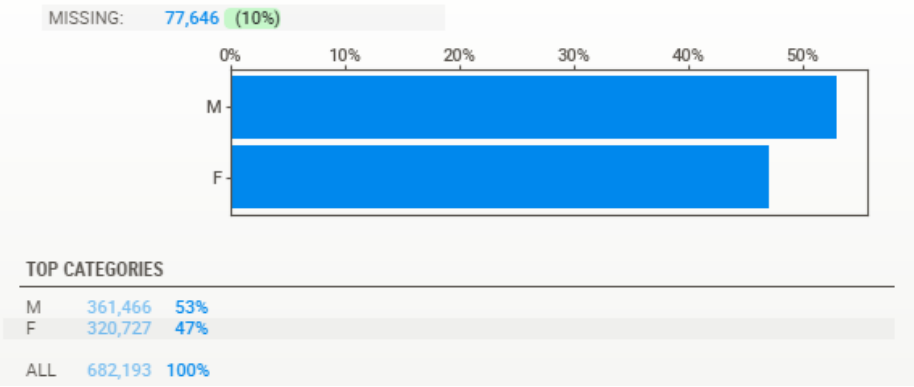

Figura 10 Distribución de género

- **Edad:** Es la edad de cada cliente y se calcula de acuerdo a la fecha de nacimiento que informa el cliente cuando ingresa a la base de datos, evidenciamos que en algunas ocasiones se diligenció mal el campo y existen muchos clientes con una edad errada, esta situación se debe corregir en la limpieza y procesamiento de los datos. En esta variable se observan algunos datos atípicos debido a errores en la recolección de la información, por ejemplo, la edad máxima registra como 2022 y la mínima como -7.938. En la figura 11 y 12 se muestra la distribución de la cantidad de clientes por edad, algunos clientes tienen el campo de edad errado, lo que genera datos atípicos, como por ejemplo que existen clientes con más de 2000 años e inferiores a -7000 años, para estos clientes y los que tienen el campo en cero o vacío se dispondrá la mediana de edades de los demás clientes. Adicional, se observa en la gráfica un pico en la edad de 54 años, esto se debe a una definición de negocio para los clientes que se registran por el canal online, ya que debido a que el campo Fecha de Nacimiento no se solicita, se está llenando de manera automática con la fecha 01/01/1970, lo que nos arroja al momento del análisis la edad de 54 años. Para no afectar los resultados del modelo, se imputa este atributo de los clientes ecommerce con la mediana de la edad de los demás clientes.

VALUES:	746,349	(98%)	MAX	2,022	RANGE	9,960
MISSING:	13,490	(2%)	95%	64	IQR	18.0
			Q3	52	STD	16.6
DISTINCT:	145	(<1%)	AVG	43	VAR	277
			MEDIAN	42		
ZEROES:	31	(<1%)	Q1	34	KURT.	72,152
			5%	25	SKEW	-136
			MIN	-7,938	SUM	32.0M

Figura 11 Distribución edades

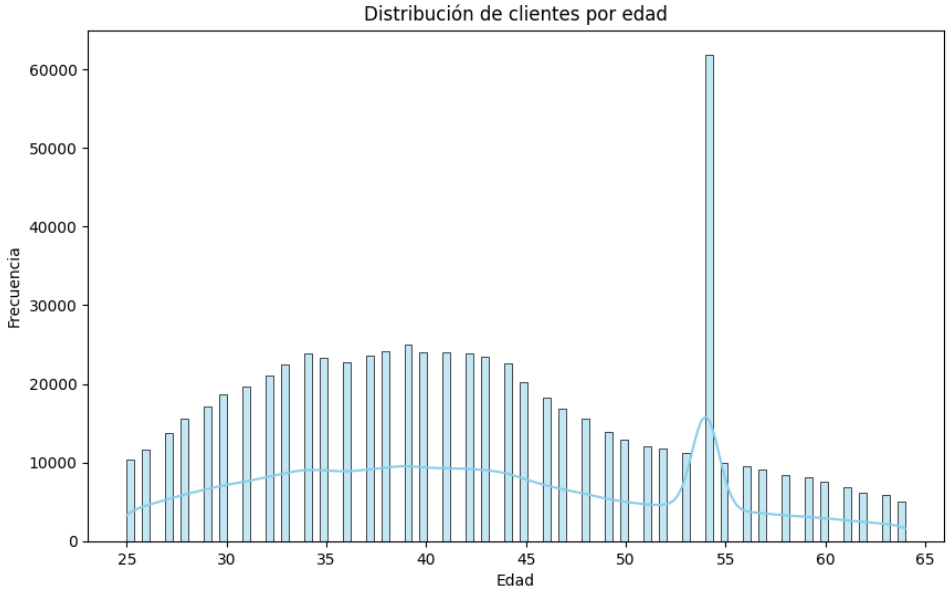

Figura 12 Distribución de clientes por edad

- **Canal:** Es el canal preferido de compra del cliente, se diferencia entre Tienda Propia, ecommerce y franquicia. El canal más usado son las tiendas propias con el 49%, seguido por franquicias con 31% (ambas son tiendas físicas) y por último el canal online con 20%. En la figura 13 se observa la distribución de porcentaje de compradores por canal de venta.

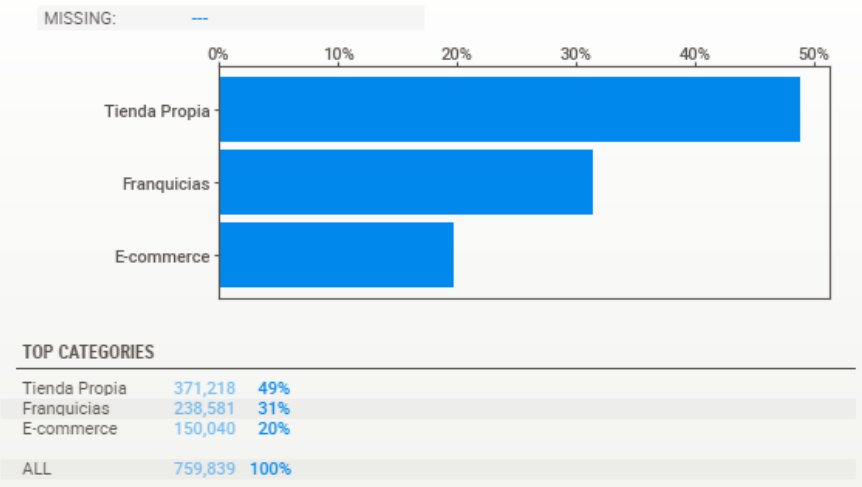

Figura 13 Distribución por canal

- **Ciudad:** Es la ciudad de la tienda preferida de cada cliente. Debemos tener en cuenta que para el canal ecommerce, por defecto se matricula Medellín como dicha Ciudad.

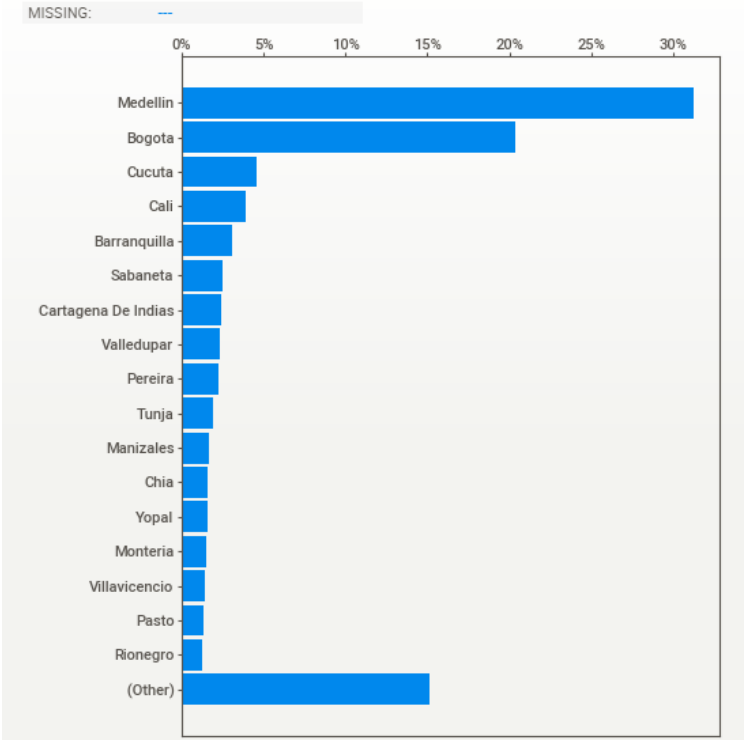

Figura 14 Distribución ciudad

- **Fecha_Reg:** Es la fecha de registro del cliente en la base de datos, este campo nos permite calcular el tiempo que lleva cada cliente en la marca. En el dataset se observa que existen clientes registrados desde hace más de 12 años (Desde el año 2012).

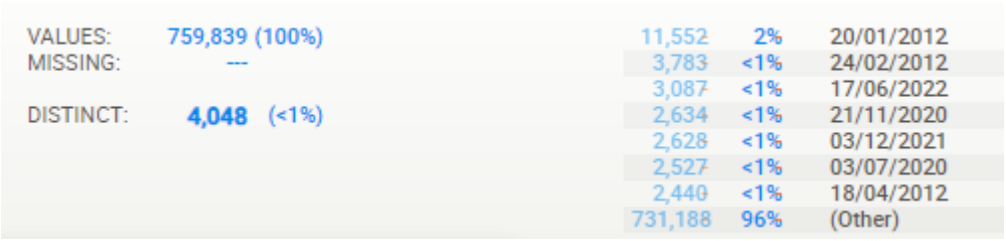

Figura 15 Fecha de registro

- **Venta:** Es la venta acumulada del cliente desde que realizó su primera compra, se tienen registros de ventas de clientes desde 2018 hasta abril de 2024. Existen algunos clientes con venta cero o negativa y eso se debe a que son devoluciones. Se observa que existen casos atípicos que tienen compras de más de 200 millones, este caso es el cliente genérico, que es el documento en el cual se registran las ventas de los clientes que no comparten su cédula para registrar dicha compra y que por regulación de la DIAN se debe registrar en un mismo número (por lo general es el 111111111).

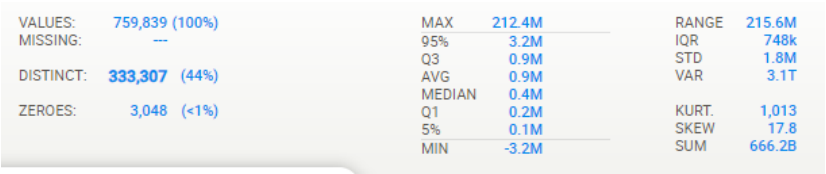

Figura 16 Valores de Venta por cliente

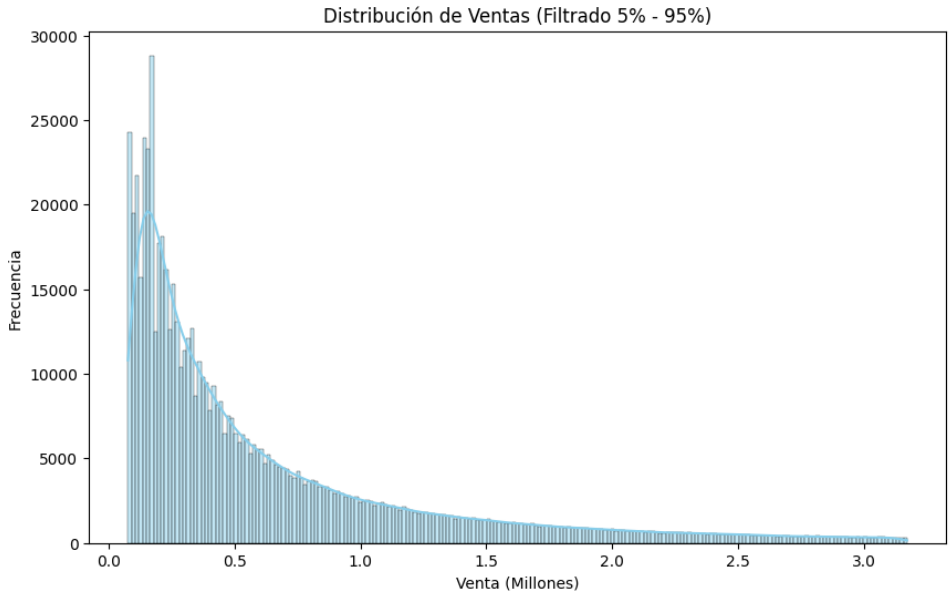

Figura 17 Distribución de ventas por cliente

- **Facturas:** Son todas las facturas que el cliente ha realizado durante su historia en la organización (Medida desde 2018). Las facturas cero corresponden a devoluciones, ya que al cliente realizar una devolución anula su factura inicial correspondiente. La mediana de facturas que han realizado los clientes en las fechas analizadas es de 2 y el promedio está en 4 facturas. Al igual que en ventas, existe un dato atípico de 637 facturas, esto corresponde al cliente genérico que por disposición de la DIAN se debe tener.

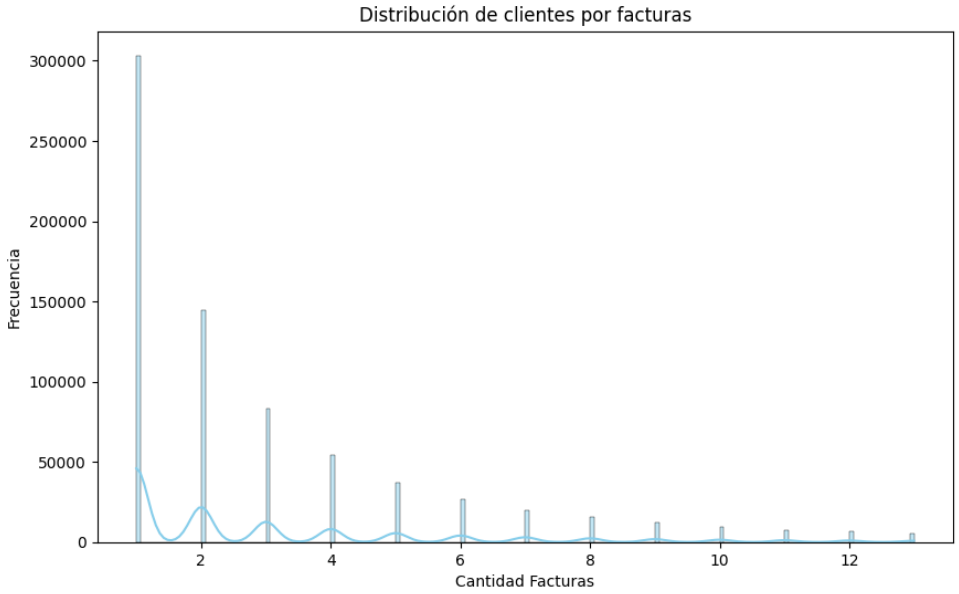

Figura 18 Distribución de clientes por facturas

VALUES:	759,839 (100%)	MAX	637	RANGE	637
MISSING:	---	95%	13	IQR	3.00
		Q3	4	STD	6.78
DISTINCT:	246 (<1%)	AVG	4	VAR	46.0
		MEDIAN	2		
ZEROES:	62 (<1%)	Q1	1	KURT.	561
		5%	1	SKEW	14.6
		MIN	0	SUM	3.0M

Figura 19 Cantidad de facturas

- **Devoluciones:** Son todos los cambios o devoluciones que el cliente ha realizado durante su historia en la organización (Medida desde 2018).Al igual que en ventas y facturas, existe un dato atípico de 453 devoluciones, esto corresponde al cliente genérico que por disposición de la DIAN se debe tener.

VALUES:	759,839 (100%)	MAX	453	RANGE	453
MISSING:	---	95%	2	IQR	0.00
		Q3	0	STD	1.49
DISTINCT:	83 (<1%)	AVG	0	VAR	2.22
		MEDIAN	0		
ZEROES:	603,867 (79%)	Q1	0	KURT.	12,702
		5%	0	SKEW	58.8
		MIN	0	SUM	286k

Figura 20 Cantidad de devoluciones

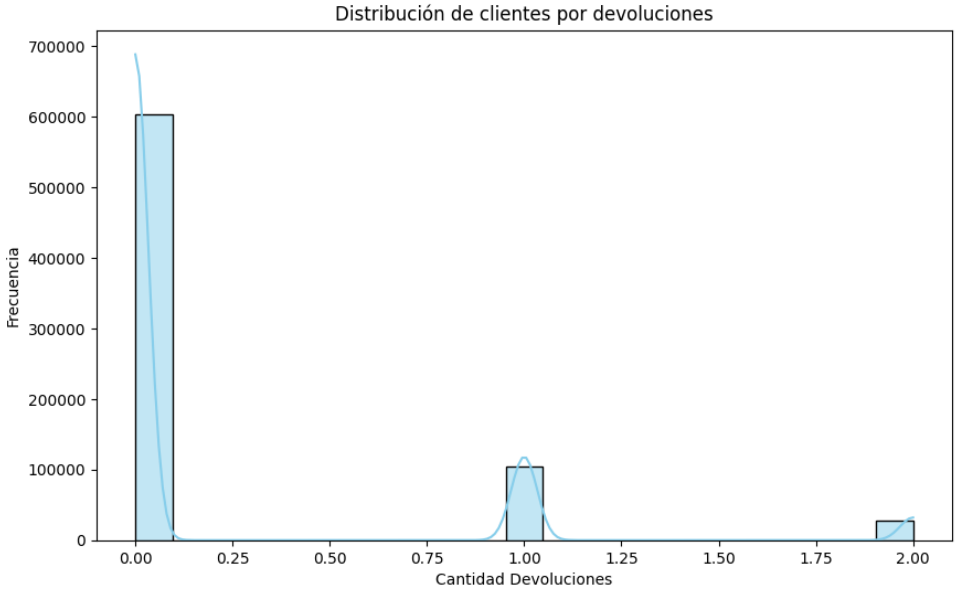

Figura 21 Distribución de devoluciones

- **Frecuencia:** Es la cantidad de veces que compra un cliente en un periodo de tiempo, para este ejercicio se toma acumulado. El cliente en promedio compra 4 veces en el periodo analizado.

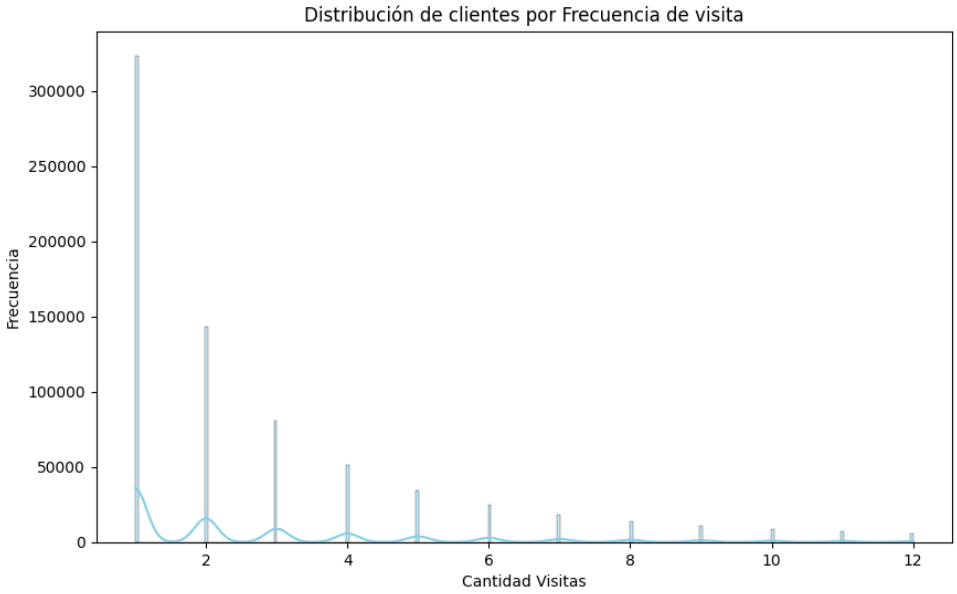

Figura 22 Distribución de cantidad de visitas (frecuencia)

VALUES:	759,839 (100%)	MAX	630	RANGE	646
MISSING:	---	95%	12	IQR	3.00
		Q3	4	STD	5.83
DISTINCT:	217 (<1%)	AVG	4	VAR	34.0
		MEDIAN	2		
ZEROES:	5,277 (<1%)	Q1	1	KURT.	546
		5%	1	SKEW	13.7
		MIN	-16	SUM	2.7M

Figura 23 Frecuencias

- **Unidades Vendidas:** Sumatoria de todas las unidades que el cliente ha comprado desde 2018. Se visualiza un outlier importante que compra 4.223 unidades, aunque el promedio de unidades de que compra cada cliente en el periodo analizado es de 8, con una mediana de 4 unidades en total.

VALUES:	759,839 (100%)	MAX	4,223	RANGE	4,223
MISSING:	---	95%	30	IQR	7.00
		Q3	9	STD	20.2
DISTINCT:	528 (<1%)	AVG	8	VAR	409
		MEDIAN	4		
ZEROES:	62 (<1%)	Q1	2	KURT.	5,458
		5%	1	SKEW	42.7
		MIN	0	SUM	6.4M

Figura 24 Distribución de unidades vendidas

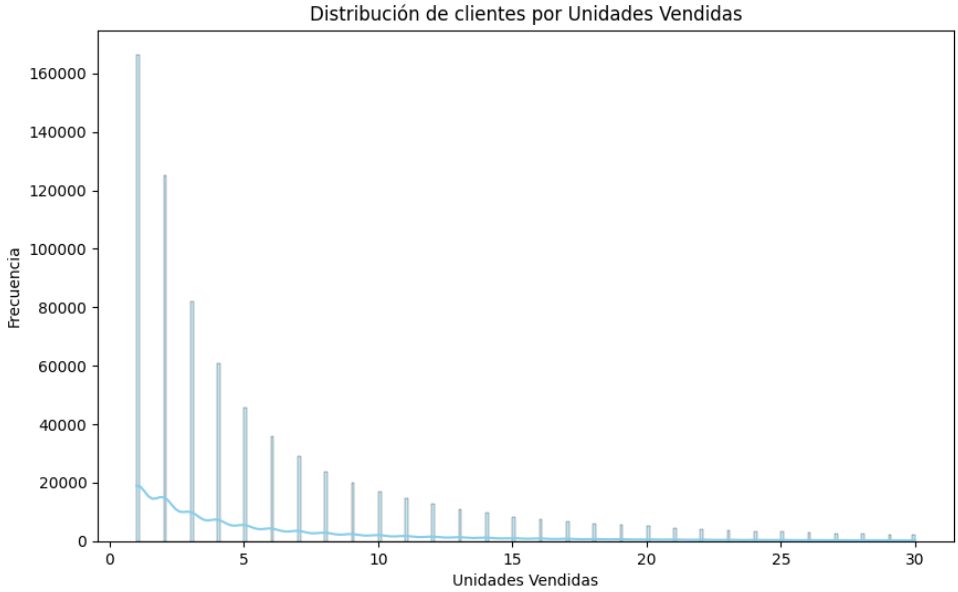

Figura 25 Distribución de clientes por Unidades Vendidas

- **Unidades Devueltas:** Sumatoria de todas las unidades que el cliente ha cambiado o devuelto desde 2018.

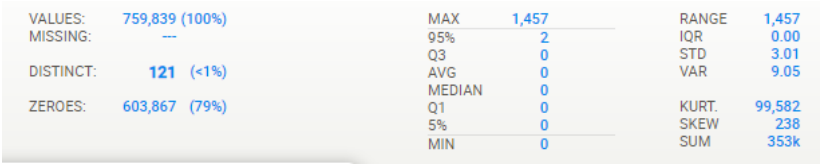

Figura 26 Distribución unidades devueltas

- **Recencia:** Son los días desde la última compra realizada por el cliente calculadas hasta el 30 de abril de 2024.

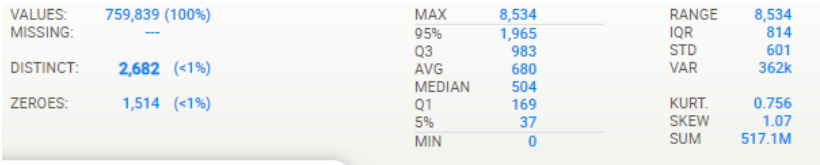

Figura 27 Distribución recencia

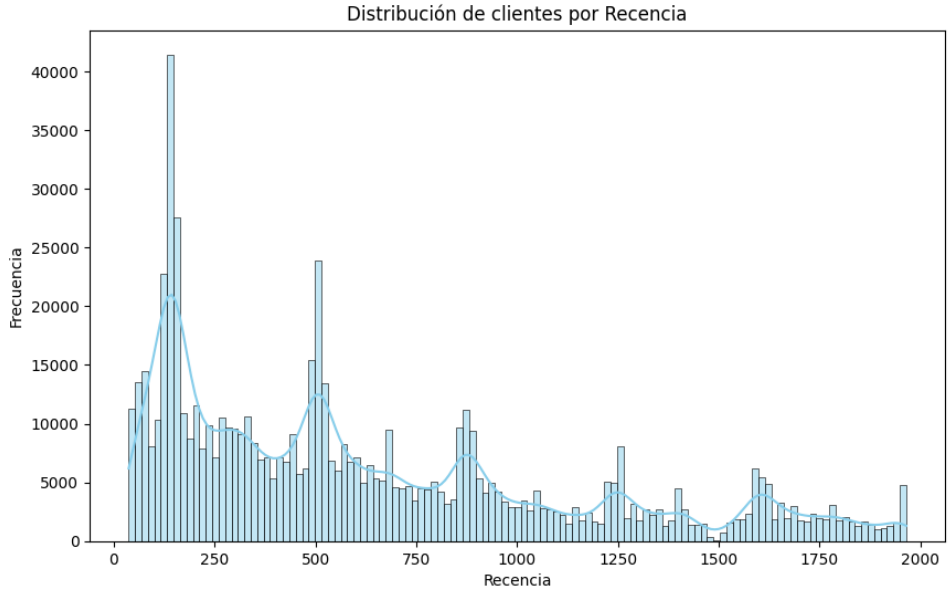

Figura 28 Distribución de clientes por Recencia

- **Cupo Tarjeta:** La marca tiene la opción de emitir tarjetas de crédito, este campo muestra el cupo aprobado para el cliente de dicha tarjeta. Se observa que el cupo promedio de los clientes es de $8.200.000 con una mediana de $6.400.000. También es importante resaltar que la gran mayoría de los clientes no tiene tarjeta (89%), por lo tanto, la variable cuenta con alto porcentaje de información faltante en sus registros.

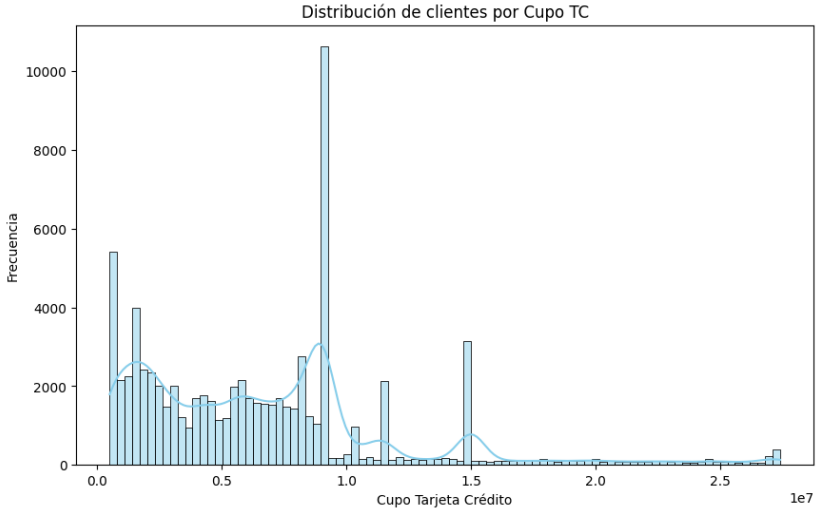

Figura 29 Distribución de clientes por Cupo TC

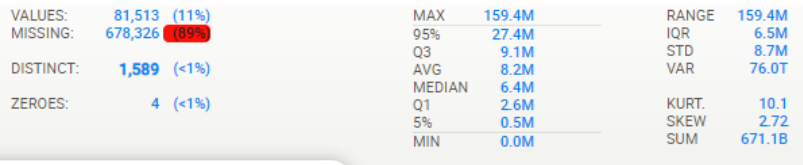

Figura 30 Cupo de tarjeta crédito

- **Estado Tarjeta:** Muestra el estado de la tarjeta, el valor 99 representa las tarjetas negadas. El 64% de los registros tiene datos perdidos y el siguiente gráfico muestra la distribución del estado de la tarjeta para el 36% restante.

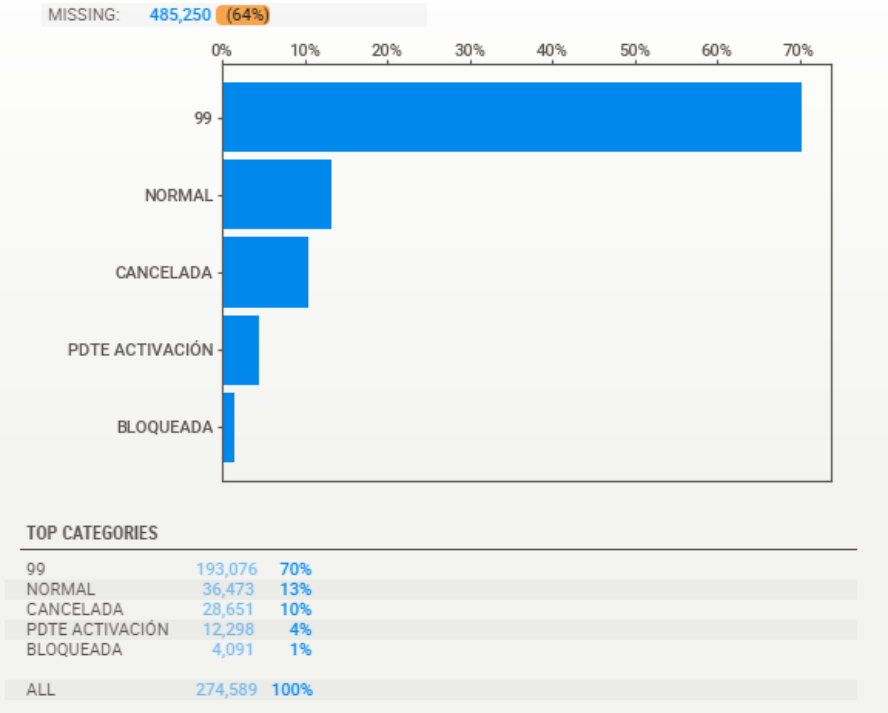

Figura 31 Estado de Tarjetas Crédito

- **Cupo Cliente:** Es el cupo de un saldo (tipo vale) que el cliente tiene con la marca para comprar sus productos y pagarlos a crédito. Se observa que el cupo promedio de los clientes es de $1.100.000 con una mediana de $900.000. También es importante resaltar que la gran mayoría de los clientes no tiene saldo-vale (93%), por lo tanto, la variable cuenta con alto porcentaje de información faltante en sus registros.

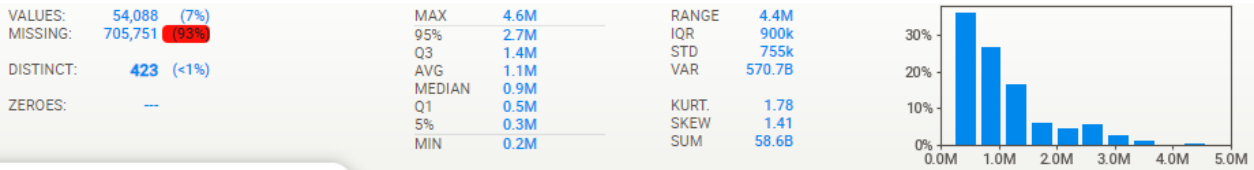

Figura 32 Cupo crédito rotativo

- **Saldo Puntos:** Es parte del programa de fidelización de la marca, permite acumular puntos por cada compra realizada para posteriormente redimirlos por productos. Se muestra el saldo acumulado al 30 de abril.

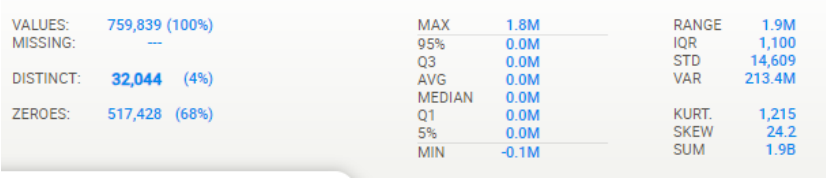

Figura 33 Saldo acumulado de puntos

- **CLTV:** Es el ciclo de vida del cliente calculado, se asigna el potencial de ingresos que podría representar cada cliente para la compañía.

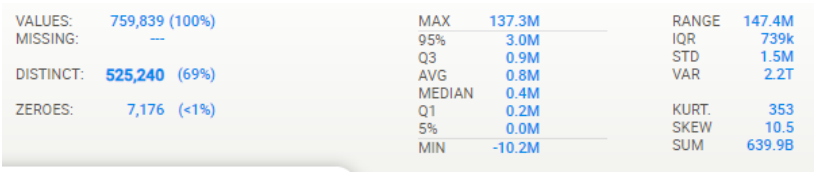

Figura 34 Customer Life Time Value

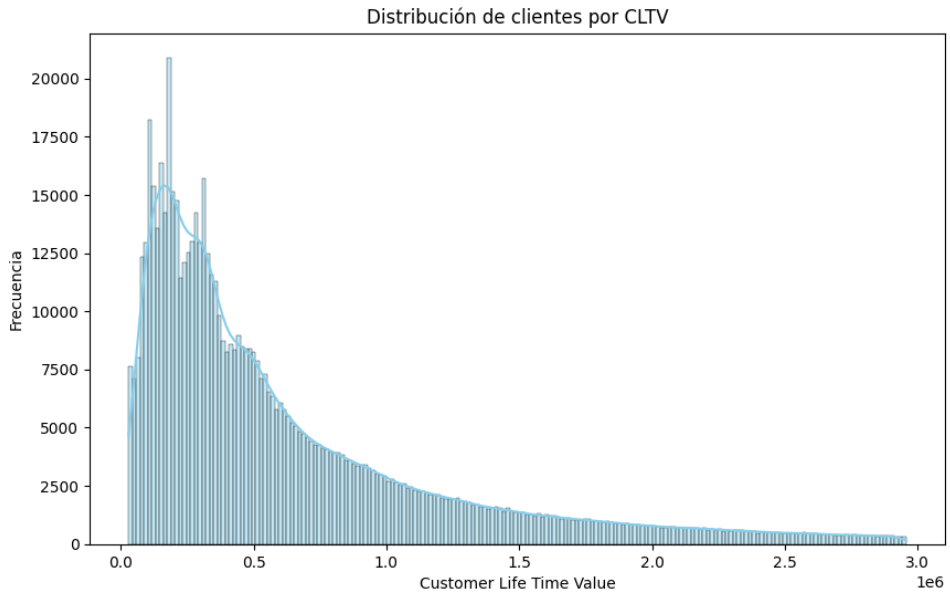

Figura 35 Distribución de clientes por CLTV

- **Frecuencia Medio de Pago:** Medimos la frecuencia con la que el cliente paga con sus diferentes medios de pago.

VALUES:	759,839 (100%)	310,527	41%	1
MISSING:	---	133,477	18%	2
		74,250	10%	3
DISTINCT:	229 (<1%)	55,753	7%	#N/D
		46,155	6%	4
		30,883	4%	5
		21,757	3%	6
		87,037	11%	(Other)

Figura 36 Frecuencia del medio de pago

- **Permanencia:** Es la cantidad de años que ha estado el cliente en la organización medida desde 2018. Más de la mitad de los clientes (52%) ha permanecido solo un año en la compañía y solo el 1% ha llegado a estar 7 años en ella. En la figura 37 se observa que el 52% de los clientes tienen 1 año de permanencia (397.550 personas), el 22% tiene permanencia de 2 años (170.487 personas), el 12% tiene 3 años de permanencia (87.467 personas), el 6% tiene permanencia de 4 años (49.141 personas) y el otro 9% tienen permanencia mayor o igual a 5 años.

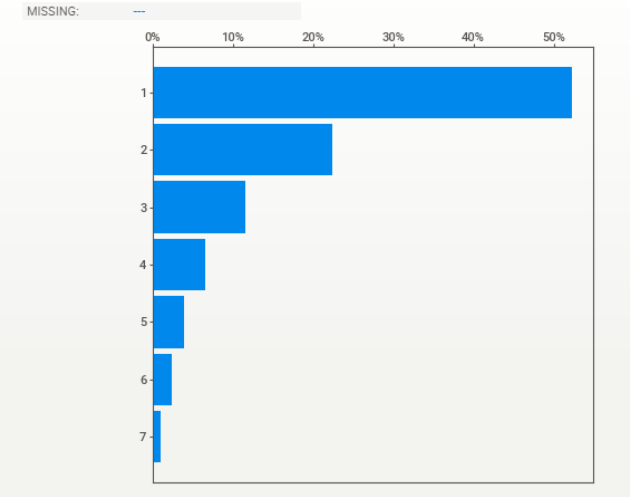

Figura 37 Permanencia del cliente en la marca

- **Tiempo en la marca:** Son los días que lleva el cliente registrado en la marca, se calculó desde la fecha de registro hasta el 30 de abril.

Figura 38 Tiempo en la marca

- **Mes UV:** Es el mes de última la última venta realizada por el cliente. Más del 30% de los clientes compró por última vez un mes de diciembre, eso nos puede dar algunas hipótesis de compras por temas estacionales.

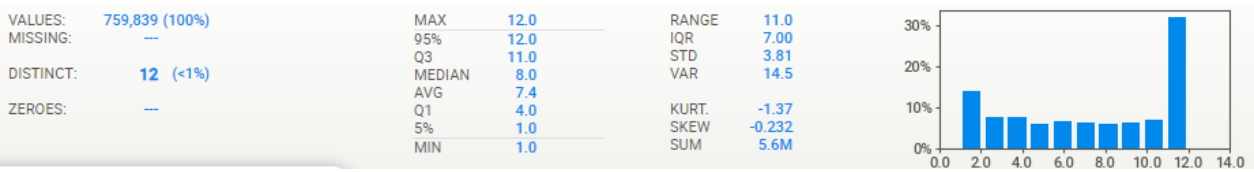

Figura 39 Mes de última venta

- **Churn:** Es la etiqueta que se va a predecir, se asigna 0 al cliente que está activo en la marca y 1 para el cliente que ya abandonó. Actualmente tenemos un 60% de clientes que ya abandonaron la marca y un 40% que aún sigue activo.

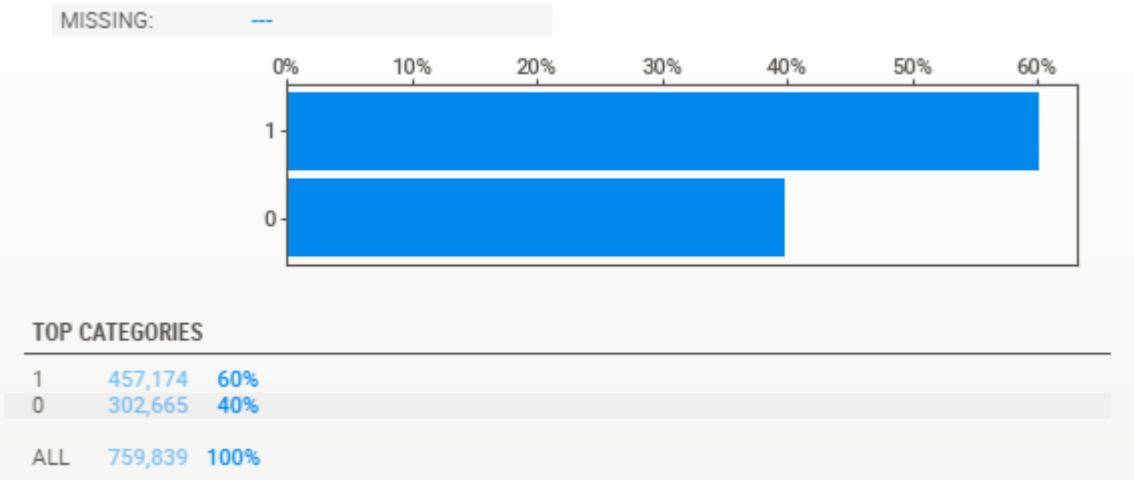

Figura 40 Etiqueta de churn

Correlaciones y asociaciones:

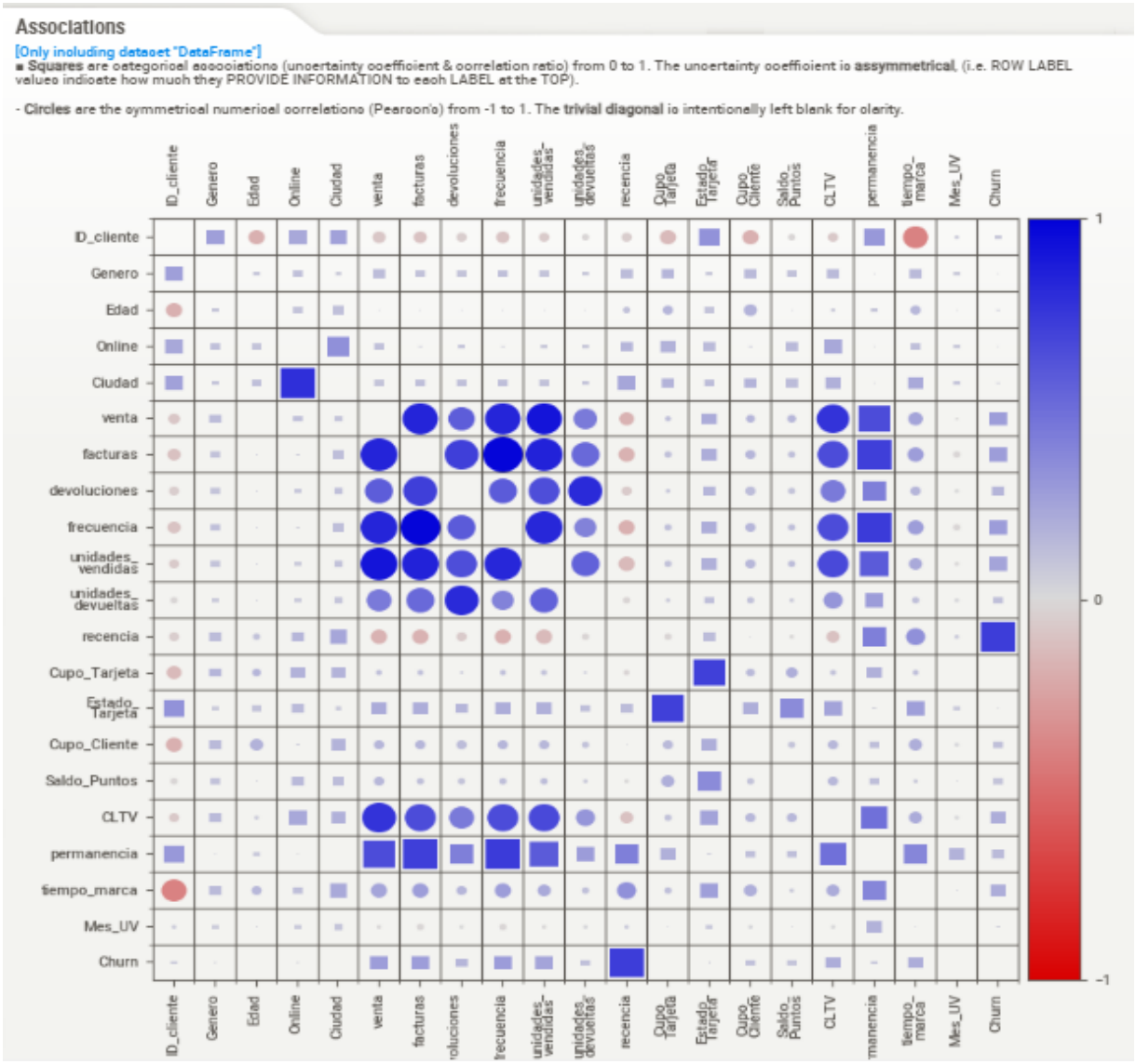

Figura 41 Correlaciones

- ■ Los cuadrados son asociaciones categóricas (coeficiente de incertidumbre y relación de correlación) de 0 a 1. El coeficiente de incertidumbre es asimétrico (es decir, los valores de ETIQUETA DE FILA indican cuánta INFORMACIÓN PROPORCIONA a cada ETIQUETA en la PARTE SUPERIOR).
- • Los círculos son las correlaciones numéricas simétricas (de Pearson) de -1 a 1. La diagonal trivial se deja intencionalmente en blanco para mayor claridad.

4.2.2 Conclusión del análisis y entendimiento de los datos

Para llegar al corpus visto en el capítulo anterior debimos explorar y comprender dentro de una gran cantidad de datos transaccionales que tiene la organización, se revisa y posteriormente se seleccionan los atributos que serían útiles en la construcción del modelo, se examinan los campos para determinar qué tipos de datos son (numéricos, strings, booleanos, fechas, etc.) y así poder asegurarnos que tuvieran la calidad adecuada, que sean datos íntegros y que tengan una estructura clara y un formato definido y puedan ser utilizados en el proyecto. Cabe resaltar que se toma la información de los clientes que tienen al menos una transacción desde enero de 2018, tenemos 22 variables explicativas (transaccionales y demográficas) y la variable a predecir, denominada Churn y que es tipo booleana, siendo 1 el cliente que abandonó (churn) y 0 el cliente que aún no abandona (no churn).

Validando los resultados del análisis exploratorio de datos, podemos identificar que tenemos algunas variables con una alta correlación, lo que nos puede generar problemas de multicolinealidad. Multicolinealidad es el nombre dado al problema que aparece cuando alguna o todas las variables independientes en una relación están altamente correlacionadas con otra. Aquí llega a ser muy difícil sino imposible, detectar sus influencias por separado y obtener estimadores razonablemente precisos de sus efectos relativos, esto es, los coeficientes de regresión parcial pueden no ser significantes, aunque exista una relación estadística entre variable dependiente y el conjunto de variables independientes. [26]

Teniendo en cuenta lo anterior, en la figura 41 de correlaciones que se encuentra en la página 39, encontramos una alta correlación entre las variables Venta con la variable Unidades Vendidas (0.91), Venta con Frecuencia (0.83) y Venta con Facturas (0.83), también encontramos que facturas se correlaciona con frecuencia (0.93) y con unidades vendidas (0.84), adicional, vemos que Unidades Devueltas se correlaciona altamente con Devoluciones (0.80). De acuerdo a los hallazgos, podríamos no tener en cuenta en el corpus de datos las unidades vendidas, la frecuencia y las unidades devueltas, esto con el fin de reducir la multicolinealidad y la dimensionalidad del mismo.

4.3 Preparación de los datos

En el proceso de preparación de los datos, hicimos una caracterización de los mismos, con el objetivo de entender su disponibilidad, ubicación (centralizada en un mismo servidor, en diversas fuentes, en diversos tipos, etc.) y así poder determinar la forma en que debíamos programar los querys o consultas en SQL o en el Cubo de datos desarrollado para gestionar los datos de los clientes en la organización. En esta preparación, tuvimos en cuenta algunos aspectos relevantes:

- Limpieza: Es el proceso de identificar y corregir o eliminar datos incorrectos, incompletos, duplicados o inconsistentes. Esto es esencial para garantizar la calidad y la fiabilidad de los resultados del análisis.
- Preprocesamiento: Se refiere a la transformación de los datos en un formato adecuado para el análisis. Esto incluye la conversión de datos categóricos a numéricos, la creación

de nuevas variables y la estandarización de los datos. Revisar principalmente los datos categóricos (Género, Ciudad, Canal) y transformarlo en variables numéricas para que los modelos pudieran fácilmente interpretarlos.

En la revisión del análisis exploratorio, encontramos que algunos campos tienen muchos datos faltantes, por ejemplo el género tiene un 10% de faltantes, el cupo tarjeta un 89% , el estado tarjeta el 89% y el cupo cliente el 93%, los campos se podrían llenar con un cero o con otro valor si así se definiera (la mediana de los datos o alguna otra estrategia de imputación) pero teniendo en cuenta que son la mayoría, se define no tener en cuenta estas variables en la modelación, ya que serán muy poco representativos y sesgarán los resultados. Para los datos categóricos que se mantendrán, se define realizar una codificación de características utilizando un esquema de codificación ordinal. Para la imputación de datos de las variables que tenían algún campo vacío, se definió lo siguiente: Para el atributo Género, se asigna “No Aplica” tanto para los NA como para los vacíos, ya que existen algunos clientes que no quieren dar su género y no es conveniente asignarle uno al azar. Para la variable Edad, se asigna la mediana a los clientes que no tienen ese campo con información. Para la variable Frecuencia Medio de Pago también se define imputarla con la mediana de los demás datos. Para las variables categóricas (Género, Canal, Ciudad) se realiza una codificación numérica con la librería sklearn a través de LabelEncoder. Al realizar la categorización en la variable género, los NA se les asigna el valor 2, a Femenino el 0 y a Masculino el 1. En el caso de canal, Ecommerce toma el valor 0, tienda propia el 2 y Franquicia el 1. Por ejemplo, para la ciudad de Medellín se asigna el 19, Bogotá el 5 y Manizales el 18.

4.4 Modelado

4.4.1 Selección de características

Teniendo en cuenta la revisión y la exploración de los datos y el análisis de correlación, se decide no tener en cuenta los siguientes atributos por su alta correlación y para evitar multicolinealidad:

- Unidades Vendidas. Se correlaciona directamente con venta (91%), facturas (84%) y frecuencia (82%).
- Frecuencia: Se correlaciona directamente con facturas (98%), venta (83%) y con unidades vendidas (82%).
- Unidades devueltas: Se correlaciona directamente con devoluciones (80%).

Adicional, tampoco tendrá los atributos que están en menos del 20% del dataset, esto quiere decir que se excluyen también los siguientes campos:

- Cupo Tarjeta
- Estado Tarjeta
- Cupo Cliente
- Saldo Puntos

En ese orden de ideas, el dataset con el que trabajaremos quedaría con las siguientes variables:

- ID_cliente
- Genero
- Edad
- Canal
- Ciudad
- Fecha_Reg
- venta
- facturas
- devoluciones
- recencia
- CLTV
- frecuencia_mediopago
- permanencia
- tiempo_marca
- Mes_UV
- Churn

La variable dependiente es el Churn, que será la característica que deseamos predecir, las demás variables seleccionadas que vemos en el listado anterior son las variables independientes que nos ayudarán a predecir el abandono o churn.

4.4.2 Enfoque de ciencia de datos y machine learning propuesto

Los modelos de machine learning supervisados juegan un papel fundamental en la predicción del riesgo de abandono de clientes en la industria del retail, ya que permiten identificar patrones complejos en grandes volúmenes de datos históricos. La capacidad de estos modelos para aprender a partir de datos etiquetados (donde se conoce el resultado, es decir, si el cliente abandonó o no) facilita la creación de predicciones precisas basadas en características como el comportamiento de compra, interacciones con la marca y factores demográficos. Una de las principales ventajas de utilizar machine learning supervisado es su capacidad para automatizar y optimizar el proceso de identificación de riesgos, generando resultados a gran escala con una velocidad y precisión que superan las técnicas tradicionales. Esto no solo permite anticipar posibles pérdidas de clientes, sino también intervenir de manera proactiva con acciones personalizadas, mejorando significativamente la eficiencia operativa y la toma de decisiones. En este proyecto se llevará a cabo un proceso iterativo entre diversos algoritmos, incluyendo Gradient Boosting, Máquinas de Vectores de Soporte, Regresión Logística, Random Forest y Redes Neuronales. El objetivo es encontrar el modelo que mejor se adapte a las necesidades y variables del negocio, arrojando las mejores métricas para una estimación precisa del riesgo de abandono. Esto permitirá implementar estrategias efectivas de retención enfocadas en los clientes identificados como de alto riesgo, maximizando su lealtad y contribuyendo al éxito a largo plazo

de la empresa.

4.4.3 Selección de técnicas de machine learning

Los modelos de machine learning que se usaron en el proyecto son: Gradient Boosting, Máquinas de Vectores de Soporte, Regresión Logística, Random Forest y Redes Neuronales Artificiales.

4.4.4 Modelado de los datos

Previamente al modelado de los datos, determinamos la distribución de los datasets en entrenamiento y validación, quedando constituidos de la siguiente manera:

1. **Conjunto de entrenamiento:** Conjunto de datos utilizado para entrenar los modelos e incluye ventas e información de los clientes desde el 2018 al 2022. Será el 70% del dataset que corresponde a 531.887 registros.
2. **Conjunto de validación:** Conjunto de datos utilizado para validar los rendimientos de los modelos e incluye ventas e información de los clientes del 2023 hasta abril de 2024. Será el 30% del dataset que corresponde a 227.952 registros.

- Se entrena cada modelo propuesto usando el conjunto de entrenamiento y se ajustan los hiperparámetros.
- Se realiza la evaluación de los modelos a través de las estimaciones realizadas utilizando las métricas definidas.
- Se comparan las métricas de evaluación para determinar el modelo que ofrece el mejor rendimiento. Dentro de las métricas consideradas en el proyecto, se consideran las siguientes: **Precisión (**$Accuracy = True\ Positive\ /\ Predicción\ Si$ **), Sensibilidad (**$Recall = True\ Positive\ /\ (True\ Positive + False\ Negative)$**),** **Especificidad (**$Especificidad = True\ Negative\ /\ \ Predicciones\ No$ **)** y **Curva ROC.**
  Las métricas de evaluación anteriores salen de la matriz de confusión.

Una vez se selecciona el mejor modelo de acuerdo con el enfoque propuesto, se evalúa su rendimiento final con nuevos datos para asegurar que los resultados son generalizables.

Para el proyecto se tuvo en cuenta la información de más de 750.000 clientes (registros) con transacciones desde el año 2018 hasta abril de 2024, pero cabe destacar que muchos de los clientes están registrados y realizando compras en la marca incluso desde el 2010. En la figura 42 se observa la cantidad de clientes que autorizan tratamiento de datos personales a partir del 2010.

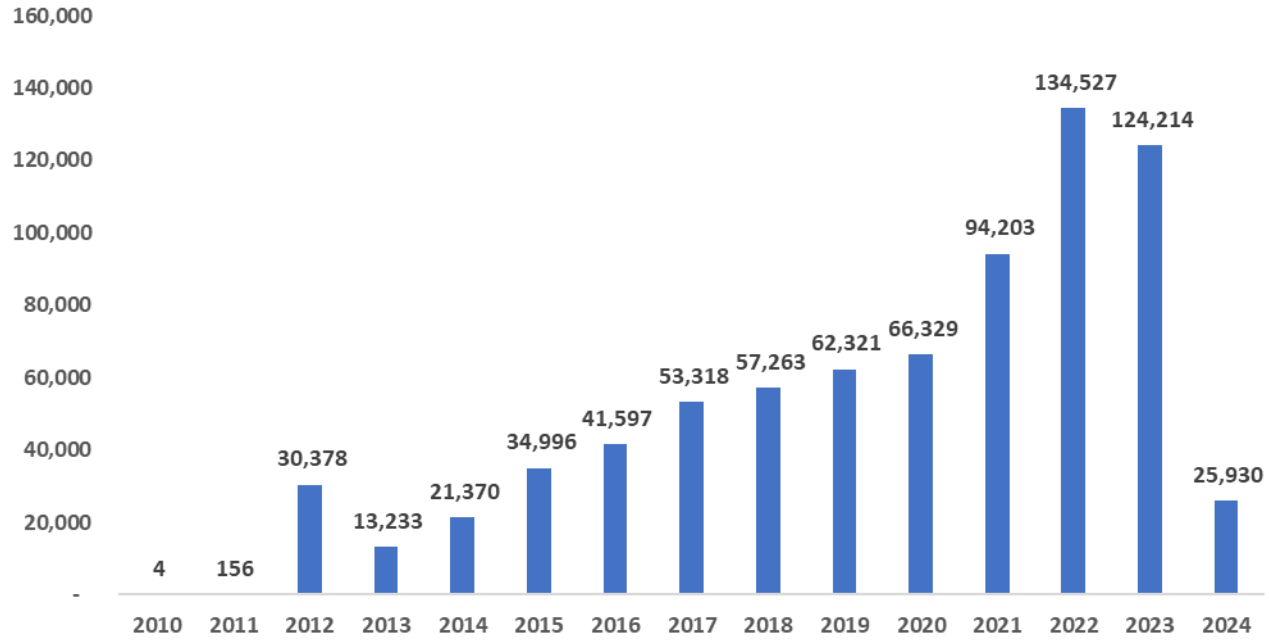

Figura 42 Cantidad de registros por año.

La implementación de machine learning para la estimación del riesgo de abandono de los clientes de la marca busca una mayor certeza en la asignación de los esfuerzos y los recursos que se usan en las estrategias de retención y de recuperación de los clientes.

4.5 Evaluación

Después de entrenar y validar los modelos propuestos, se presentan los resultados de los modelos evaluados y las métricas de rendimiento consideradas dentro del proyecto.

Tabla 2 Métricas de evaluación de los modelos

Métrica	Regresión Logística	Random Forest	X G Boost	SVM	ANN
Accuracy	0,7579	0,8586	0,8618	0,7686	0,8567
Recall (Sensibilidad)	0,8242	0,869	0,8835	0,7269	0,8828
Especificidad	0,6915	0,8482	0,8401	0,8109	0,8305
Curva ROC - AUC	0,7932	0,9462	0,9485	0,8102	0,94
Tiempo promedio ejecución (Seg)	0,7417	87,4815	7,0558	2700	1600,9444

Revisando los resultados, vemos que todos los modelos tienen muy buen performance, accuracys superiores a los 70 puntos y sensibilidades superiores al 80.

El modelo XG Boost presentó los mejores resultados, pero por muy poco sobre los demás analizados. Tuvo una precisión del 86.18% y una sensibilidad, que para nuestro caso será la métrica más importante a tener en cuenta, de un 88.35%. A continuación, presentamos la gráfica de la Curva ROC y al área bajo la curva con un puntaje de 94.85%

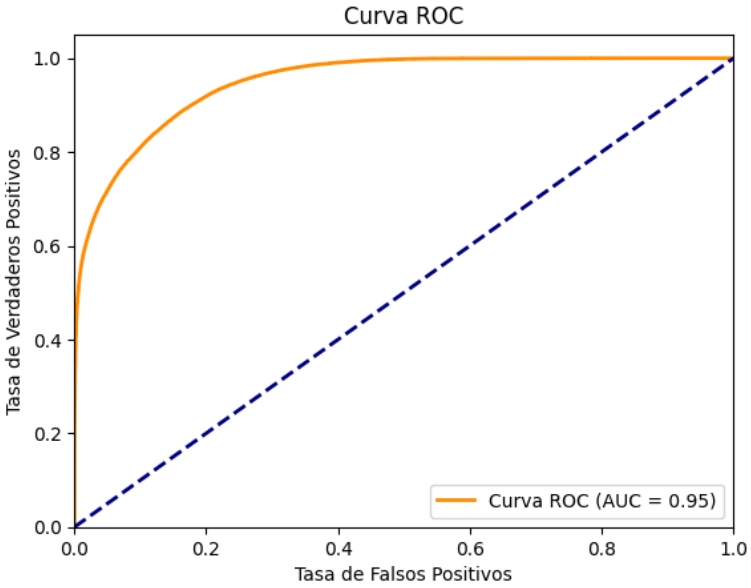

Figura 43 Curva ROC XG Boost

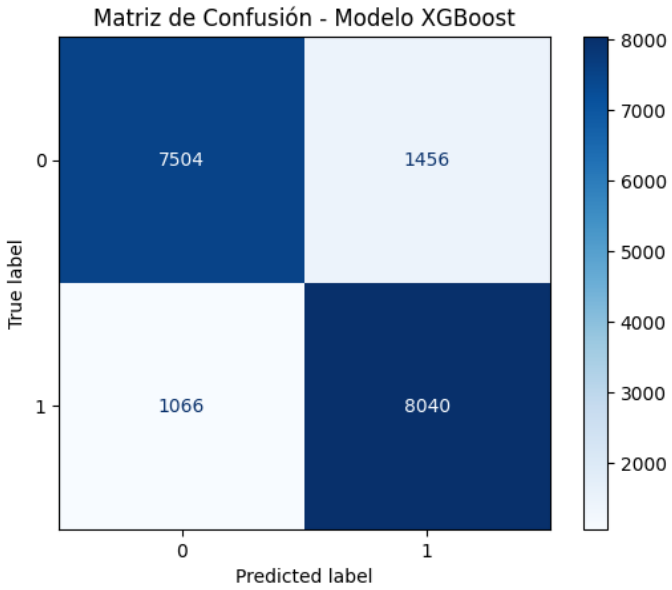

Figura 44 Matriz de Confusión - Modelo XGBoost

En la figura 44 se observa la matriz de confusión para el modelo XG Boost, donde nos muestra que tenemos 75.040 clientes que el modelo estimó que no abandonarían y no abandonaron (Verdadero positivo), mientras que estimó también 80.400 clientes que abandonarían y efectivamente abandonaron (Verdadero negativo). Se debe tener en cuenta que la matriz de confusión se crea sobre el set de validación que es el 30% del dataset balanceado que está compuesto por 602.180 clientes.

El modelo Random Forest también tuvo un muy buen desempeño, con un accuracy de 85.86% y una sensibilidad de 86.9%. Anexamos la curva ROC y las variables más importantes para el modelo. Para resaltar, se observa en la figura 45 que el tiempo que un cliente ha pasado con la marca es un factor clave para predecir la variable objetivo (abandono).

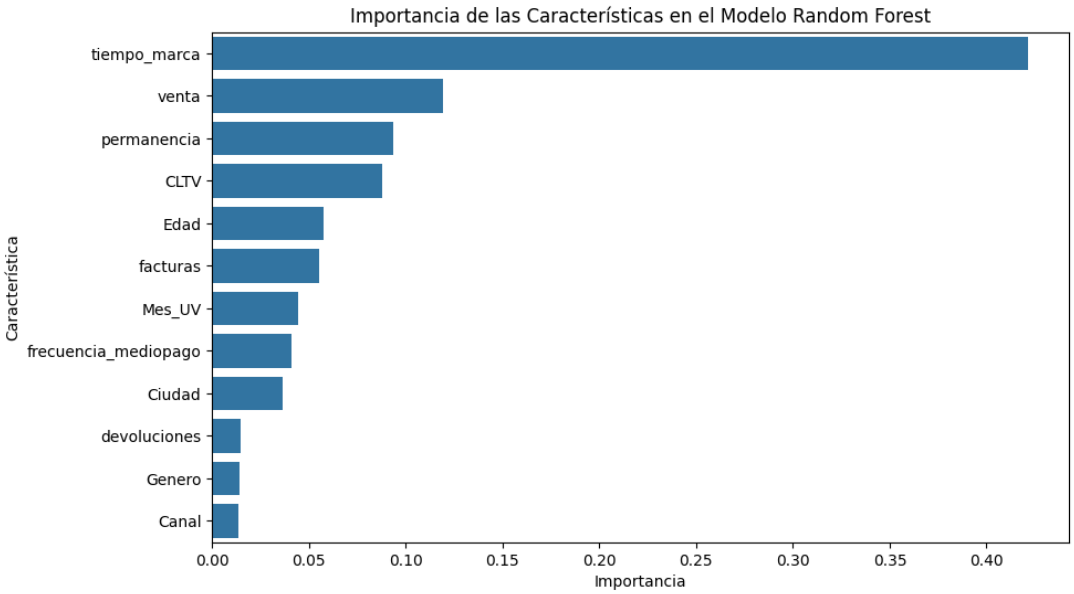

Figura 45 Importancia de las Características en el Modelo Random Forest

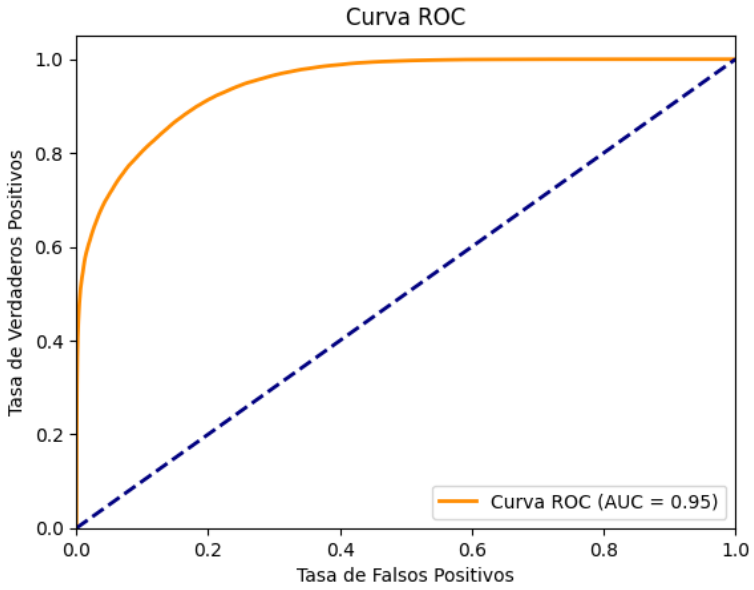

Figura 46 Curva ROC Random Forest

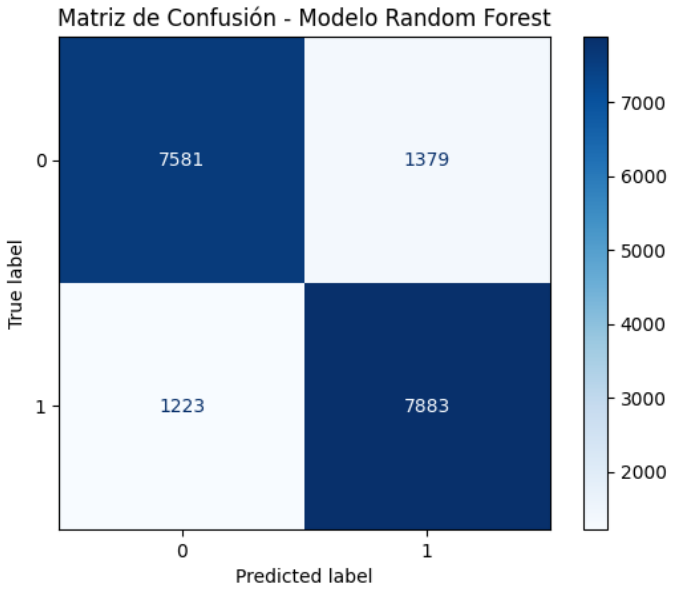

Figura 47 Matriz de Confusión - Modelo Random Forest

En la figura 47 se observa la matriz de confusión para el modelo random forest, donde nos muestra que tenemos 75.810 clientes que el modelo estimó que no abandonarían y no abandonaron (Verdadero positivo), mientras que estimó también 78.830 clientes que abandonarían y efectivamente abandonaron (Verdadero negativo). Se debe tener en cuenta que la matriz de confusión se crea sobre el set de validación que es el 30% del dataset balanceado que está compuesto por 602.180 clientes.

El modelo SVM tuvo un muy buen desempeño, con un accuracy de 76.86% y una sensibilidad de 72.69%. Anexamos la curva ROC y la gráfica de importancia de las variables.
De acuerdo a la figura 46, vemos la relación entre la importancia de las variables y el riesgo de abandono, se observa una relación negativa, es decir, que a medida que aumenta la permanencia y la venta, el riesgo de abandono disminuye.

En la figura 48 se observa la matriz de confusión para el modelo SVM, donde nos muestra que tenemos 72.180 clientes que el modelo estimó que no abandonarían y no abandonaron (Verdadero positivo), mientras que estimó también 66.550G clientes que abandonarían y efectivamente abandonaron (Verdadero negativo). Se debe tener en cuenta que la matriz de confusión se crea sobre el set de validación que es el 30% del dataset balanceado que está compuesto por 602.180 clientes.

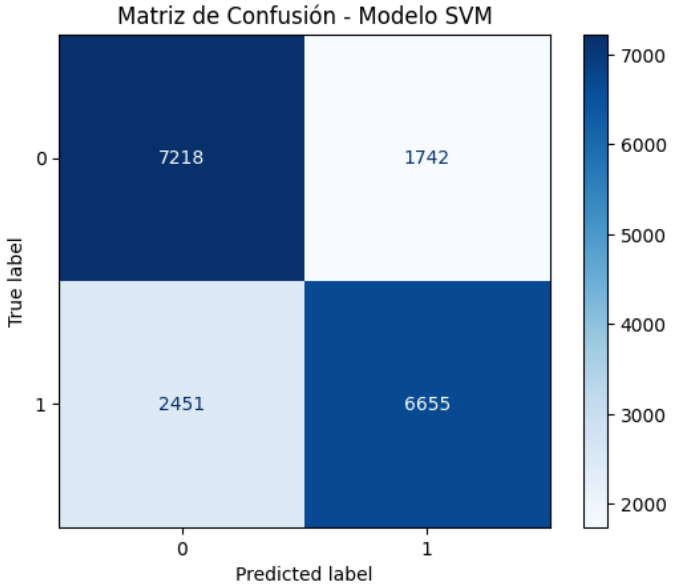

Figura 48 Matriz de Confusión - Modelo SVM

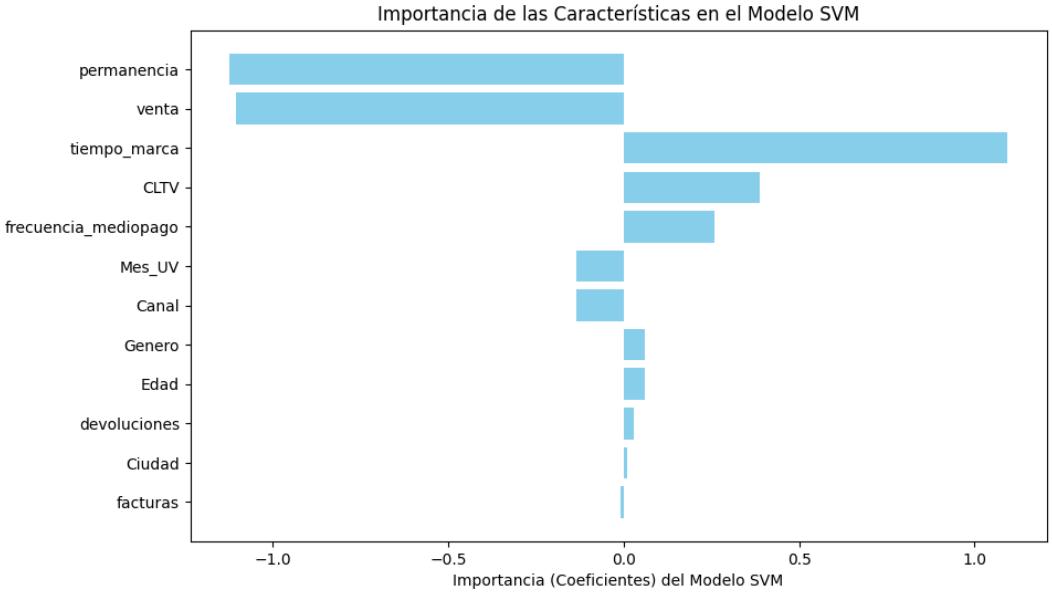

Figura 49 Importancia de las Características en el Modelo SVM

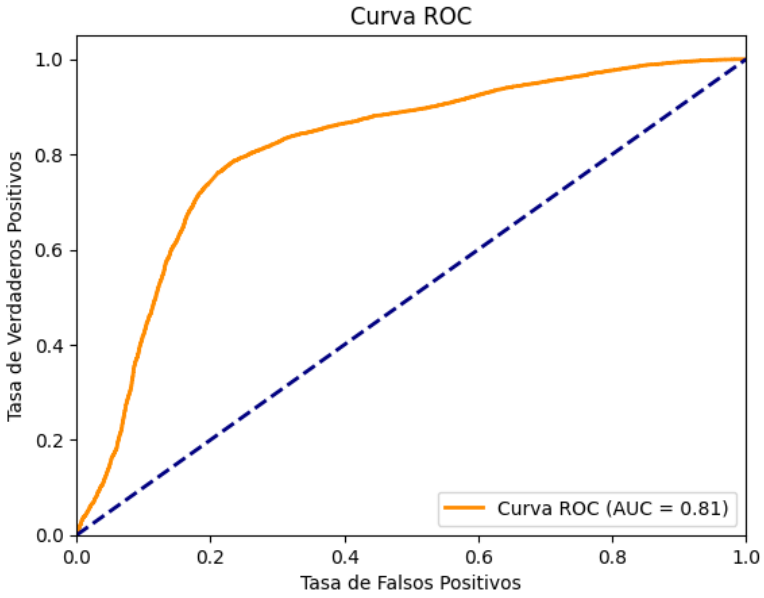

Figura 50 Curva ROC Support Vector Machine

Con un desempeño sobre saliente está la red neuronal artificial (ANN)que tuvo un accuracy de 85.67%, una sensibilidad de 88.28% pero un tiempo de ejecución bastante elevado a comparación de los demás (1600 segundos – 26.6 minutos). Anexamos la curva ROC.

En la figura 51 se observa la matriz de confusión para el modelo ANN, donde nos muestra que tenemos 73.650 clientes que el modelo estimó que no abandonarían y no abandonaron (Verdadero positivo), mientras que estimó también 79.660 clientes que abandonarían y efectivamente abandonaron (Verdadero negativo). Se debe tener en cuenta que la matriz de confusión se crea sobre el set de validación que es el 30% del dataset balanceado que está compuesto por 602.180 clientes.

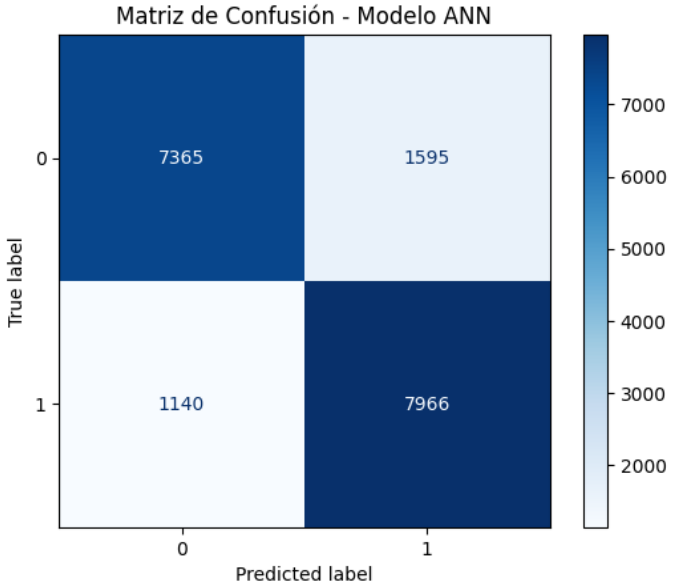

Figura 51 Matriz de Confusión - Modelo ANN

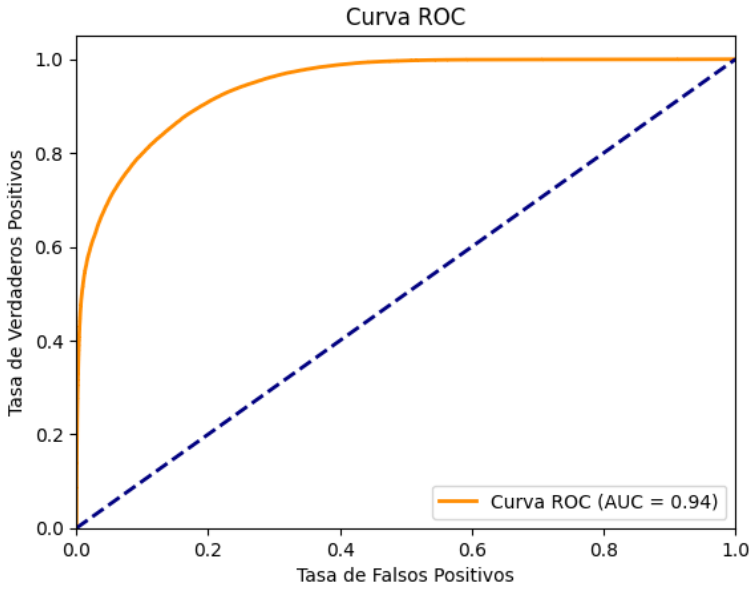

Figura 52 Curva ROC Redes Neuronales Artificiales - ANN

La regresión logística tuvo un desempeño significativo, pero no estuvo a la altura de los demás modelos, ya que su curva ROC está en un 80%.

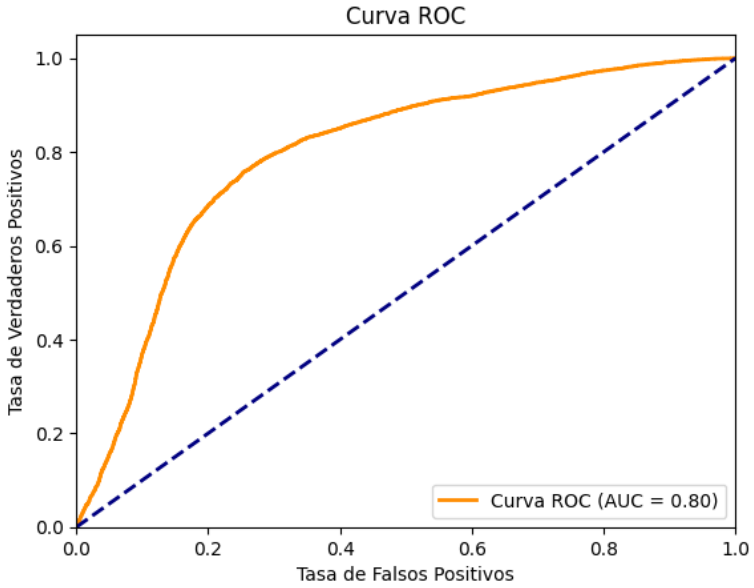

Figura 53 Curva ROC Regresión Logística

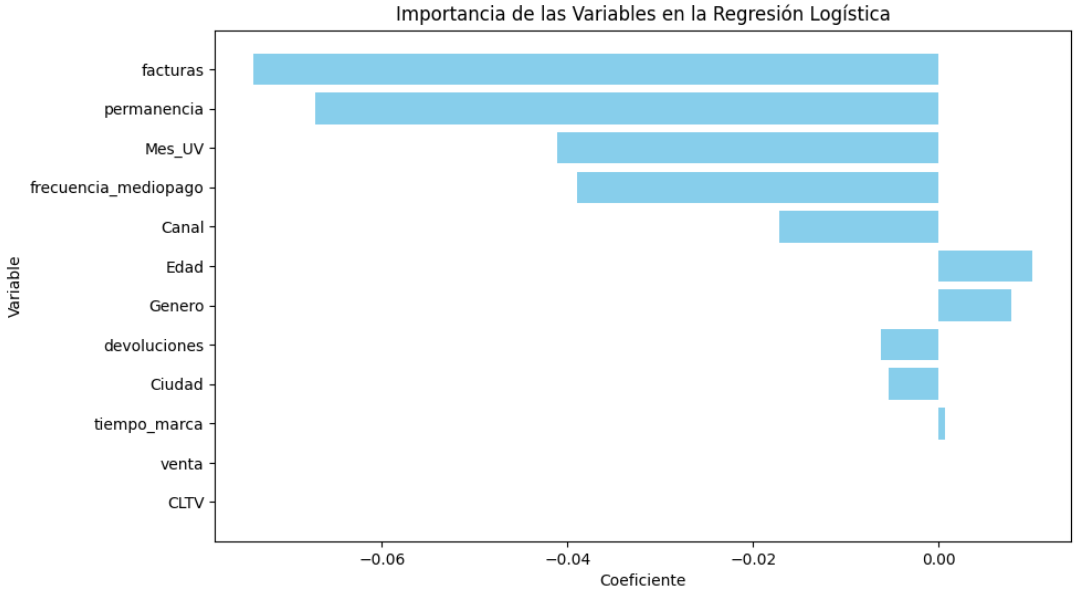

Figura 54 Importancia de las Variables en la Regresión Logística

En la figura 55 se observa la matriz de confusión para el modelo de regresión logística, donde nos muestra que tenemos 66.970 clientes que el modelo estimó que no abandonarían y no abandonaron (Verdadero positivo), mientras que estimó también 69.130 clientes que abandonarían y efectivamente abandonaron (Verdadero negativo). Se debe tener en cuenta que la matriz de confusión se crea sobre el set de validación que es el 30% del dataset balanceado que está compuesto por 602.180 clientes.

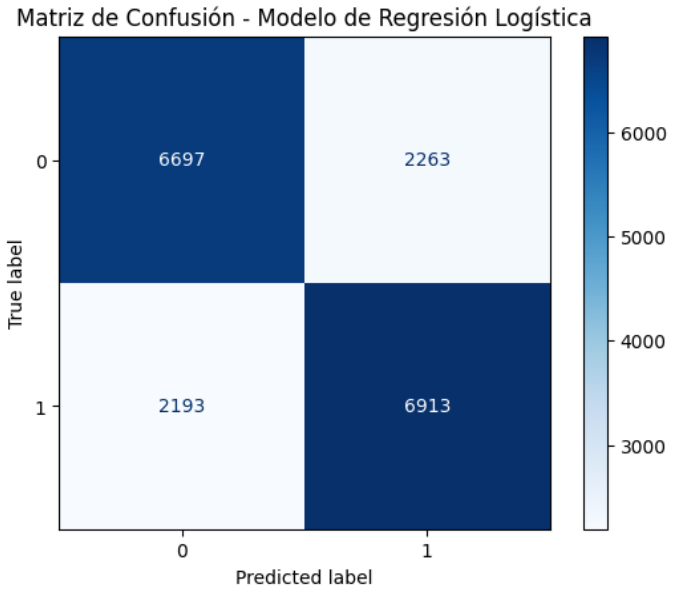

Figura 55 Matriz de Confusión - Modelo de Regresión Logística

De acuerdo a la figura 56, vemos la relación entre la importancia de las variables y el riesgo de abandono, se observa una relación negativa, es decir, que a medida que aumenta las facturas, la permanencia el riesgo de abandono disminuye. Las barras más largas representan variables con un mayor impacto en la predicción. En esta gráfica, variables como **facturas**, **permanencia** y **Mes_UV** tienen los coeficientes negativos de mayor magnitud, lo cual sugiere que son las más influyentes en el modelo.

Considerando los resultados de los diferentes modelos, se puede concluir que cualquier modelo podría ser útil para la predicción del abandono, ya que todos tienen unos excelentes desempeños (basados en sus métricas de evaluación) y la elección dependerá de los objetivos específicos de la empresa, como recomendación, se podrían seguir mapeando todas las alternativas e irlos alternando teniendo en cuenta las variables elegidas y la capacidad computacional, adicional, se debe garantizar su correcta administración y actualización con el paso del tiempo.
En cuanto a la importancia de las variables tenidas en cuenta dentro de la investigación, utilizadas en el entrenamiento y evaluación de los modelos, es clave resaltar que la relevancia de estas variables fue analizada mediante librerías de scikit-learn y seaborn que permite evidenciar la importancia y el impacto de cada variable en las predicciones del modelo. En la figura 56 se muestra la importancia de las características para el modelo elegido (XG Boost).

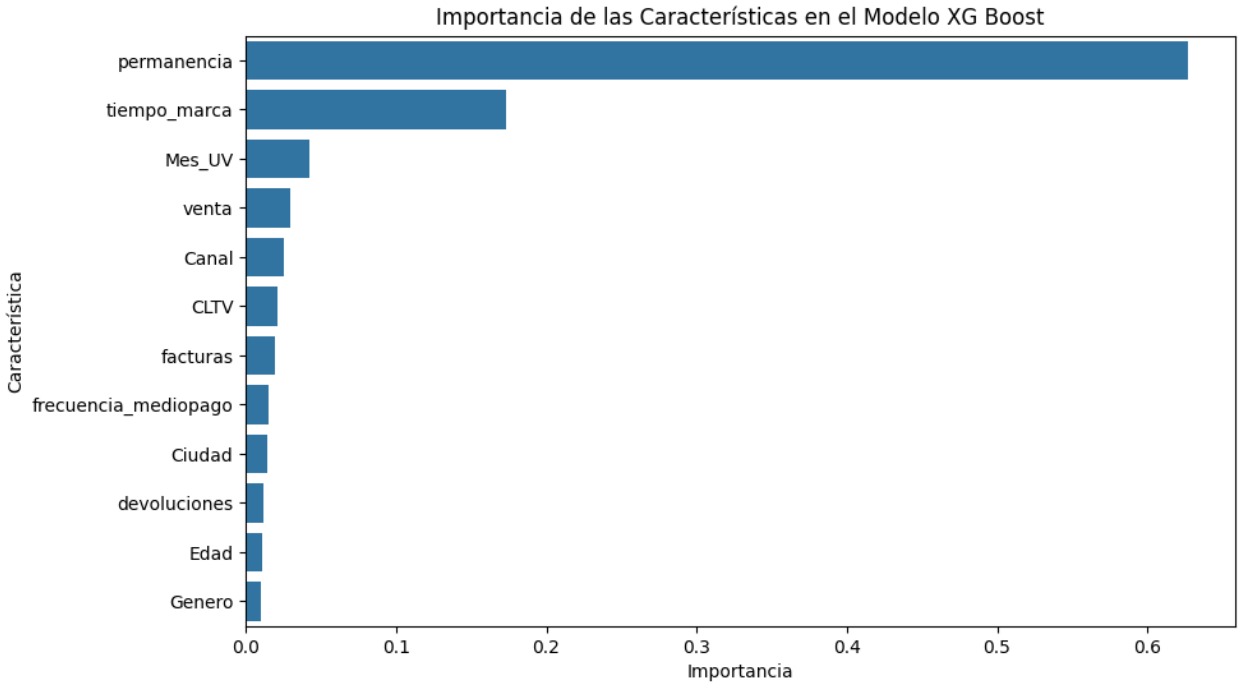

Figura 56 Importancia de las características

En la gráfica anterior visualizamos que para el modelo XG Boost, la variable más importante es la permanencia con casi 0.6, seguido del tiempo del cliente en la marca con algo más de 0.2. Esto sugiere que la duración de la relación con un cliente y el tiempo que lleva utilizando una determinada marca son factores muy influyentes en el resultado que el modelo está tratando de predecir. En el otro extremo, características como "Genero" y "Edad" parecen tener una influencia mucho menor en las predicciones. Esto no significa que estas características sean irrelevantes, sino que su impacto en el resultado final es menor en comparación con otras variables.

Con la ayuda del método SHAP graficamos la relación entre las variables “Permanencia” y “Tiempo Marca” y encontramos lo siguiente:

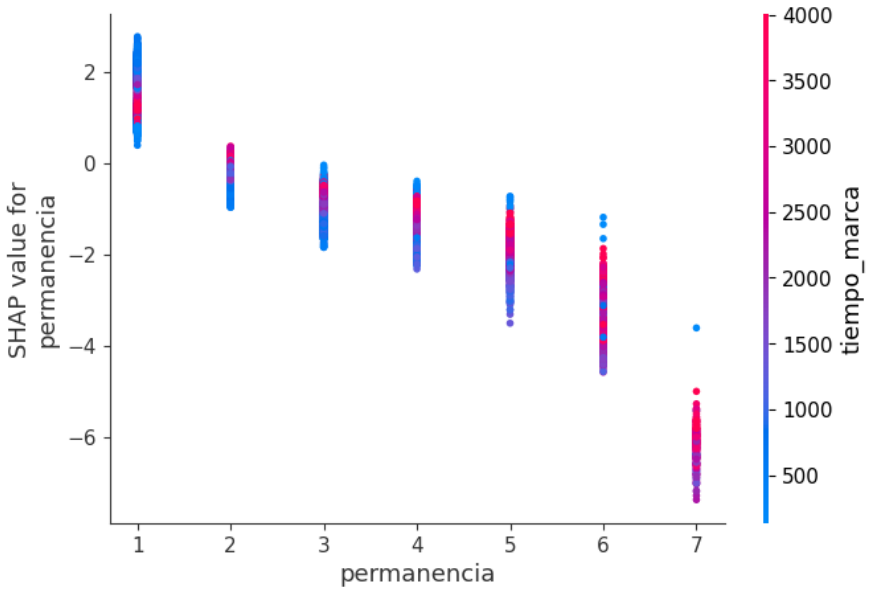

Figura 57 Relación Permanencia – Tiempo Marca

4.5.1 Relación Principal:

- La dispersión de puntos a lo largo del eje x (permanencia) nos indica que, en general, a medida que aumenta la "permanencia", los valores de SHAP también tienden a disminuir. Esto sugiere una relación negativa entre "permanencia" y la predicción del modelo). Es decir, a mayor permanencia, menor es el valor predicho. En síntesis, entre mayor permanencia tenga el cliente en la marca (lleve más años comprando), es mucho menor la probabilidad del cliente de abandonarla.

4.5.2 Efecto de la Interacción con "tiempo_marca":

- La coloración de los puntos representa los diferentes valores de "tiempo_marca". Los puntos azules corresponden a valores bajos de "tiempo_marca", mientras que los puntos rojos corresponden a valores altos.
- Se observa que para valores bajos de "permanencia" (hacia la izquierda del gráfico), la variabilidad en los valores de SHAP es mayor y parece haber una mayor influencia de "tiempo_marca". Es decir, para clientes con poca permanencia, el "tiempo_marca" juega un papel más importante en la predicción.
- A medida que aumenta la "permanencia", la influencia de "tiempo_marca" parece disminuir y los valores de SHAP se vuelven más consistentes. Esto sugiere que, para clientes con alta permanencia, otros factores además del "tiempo_marca" pueden estar influyendo más en la predicción.

4.6 Implementación

Después de la realización de los modelos y sus respectivas evaluaciones, se define que se llevará a producción e implementación el XG Boost, debido a que fue el mejor modelo en cuanto a las métricas más relevantes (Precisión y sensibilidad) y además toma poco tiempo entrenarlo. Para implementarlo debemos tener en cuenta lo siguiente.

4.6.1 Preparación del entorno: Servidor on-premise dentro de la oficina en Medellín con capacidad dedicada para el modelo, configuración correcta de Python con el entorno actualizado con las bibliotecas correspondientes y el flujo de datos asegurado desde la fuente.

4.6.2 Integración con la herramienta OLAP: El resultado de la predicción del riesgo de abandono se integrará directamente con la base de datos relacional en un nuevo campo que se llama Riesgo Fuga y que permita su visualización dentro del cubo de datos. El nuevo campo Riesgo Fuga tendrá unos cálculos definidos de la siguiente forma: Alto con probabilidad mayor al 80%, Medio que sea mayor a 40% y menor a 80% y bajo que sea menor a 40%. La definición de los rangos se hizo de acuerdo a un criterio corporativo después de una discusión con el equipo de mercadeo. El nuevo campo Riesgo Fuga se integrará en la Dimensión Cliente del cubo de datos OLAP (Ver Figura 60)

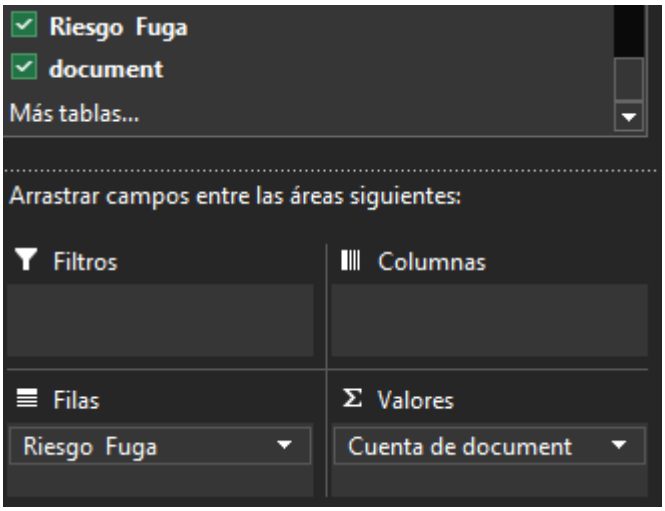

Figura 58 Vista de predicción en el cubo

Etiquetas de fila	Cuenta de document
Alto	150,825
Bajo	90,315
Medio	283
Total general	**241,423**

Figura 59 Vista de predicción en el cubo (cantidades)

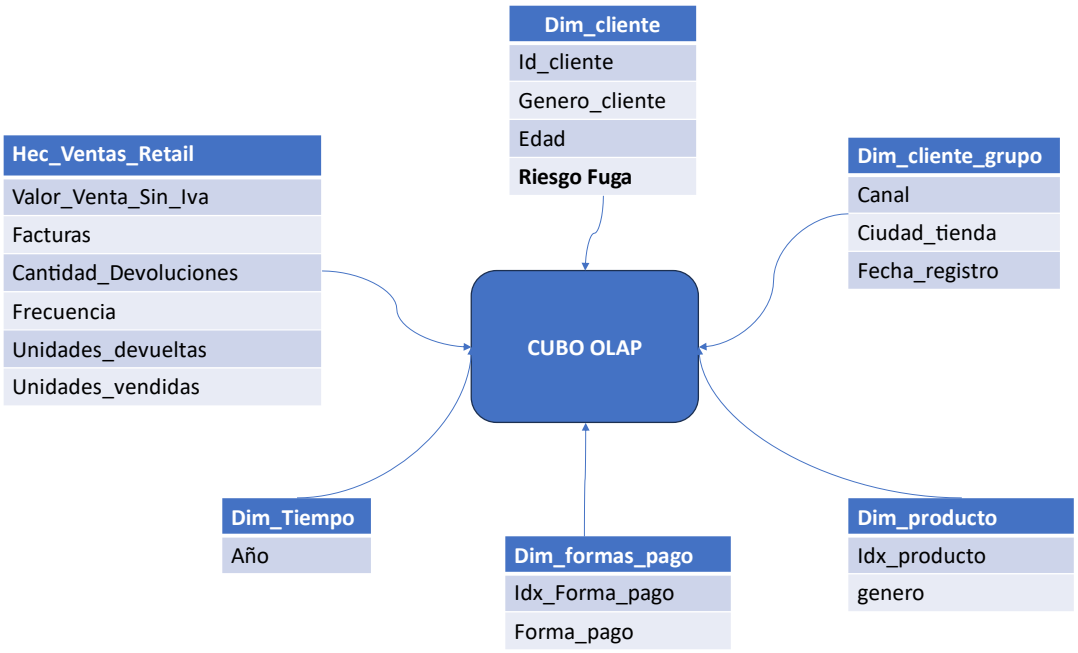

Figura 60 Diagrama del cubo con el nuevo campo Riesgo Fuga

4.6.3 Monitoreo y Mantenimiento: Se monitorea el rendimiento del modelo cada mes en producción para realizar los ajustes necesarios (cuando aplica). Adicional, cada mes se valida aleatoriamente un grupo de clientes que el modelo haya predicho y validamos su estado, si lo predijo como riesgo alto, validamos si efectivamente abandonó y si lo predijo como riesgo bajo, validamos que aun esté activo en la compañía, además, se reentrena el modelo con nuevos datos cada 3 meses.

Descripción del diagrama propuesto:

1. **Ingesta de Datos:** Los datos de los clientes se recolectan de diversas fuentes (CRM, transacciones, etc.) y se almacenan en un data warehouse.
2. **Preprocesamiento:** Los datos se limpian, transforman y preparan para ser utilizados por el modelo de XGBoost.
3. **Entrenamiento del Modelo:** El modelo se entrena con los datos históricos y se evalúa su rendimiento.
4. **Implementación en Producción:** El modelo se despliega en un entorno de producción y se conecta a la herramienta OLAP.
5. **Cubo OLAP:** Los datos del modelo se cargan en un cubo OLAP, creando dimensiones y medidas relevantes.
6. **Consulta y Visualización:** Los usuarios finales utilizan la herramienta OLAP para consultar

las predicciones, crear informes y visualizar los resultados.

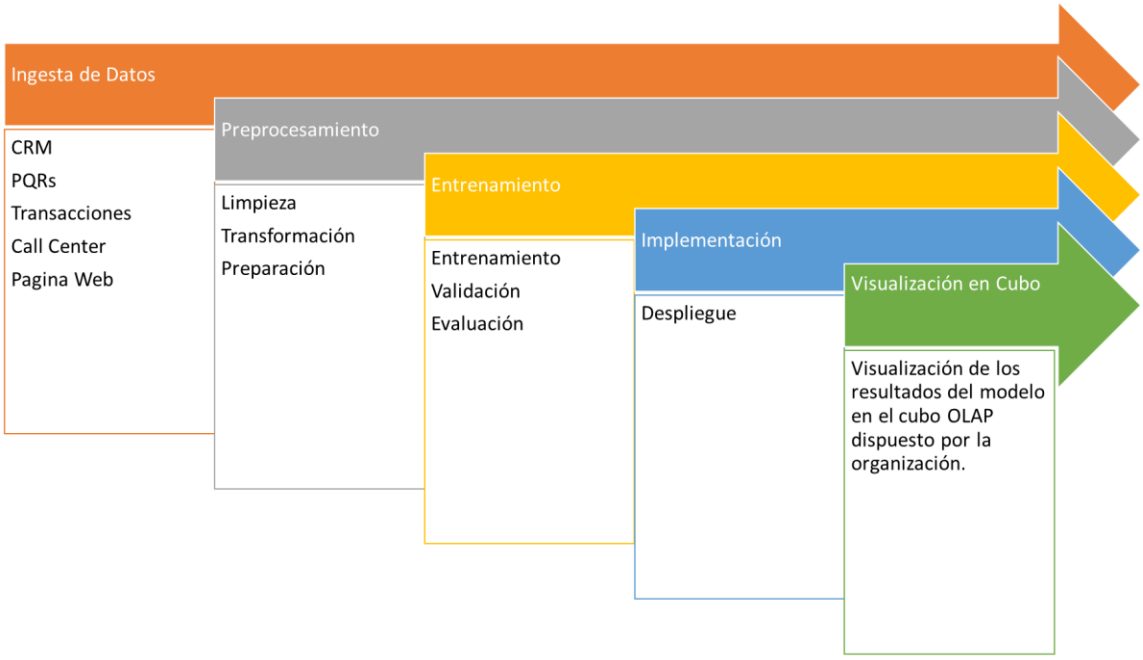

Figura 61 Diagrama propuesto

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1 CONCLUSIONES

El desarrollo de un modelo predictivo para estimar el riesgo de abandono de clientes en una empresa de retail de moda, utilizando técnicas de machine learning, ha demostrado ser un enfoque sumamente efectivo. A lo largo del proyecto, se probaron diversos algoritmos, entre ellos XGBoost, Random Forest, Support Vector Machines (SVM) y redes neuronales artificiales (ANN), y los resultados arrojaron métricas de desempeño sobresalientes. El modelo XGBoost fue el que mostró un mejor desempeño, con una precisión del 86.18% y una sensibilidad del 88.35%, lo que lo convierte en la mejor opción para estimar el riesgo de abandono de clientes.

Una de las principales conclusiones del trabajo es que los modelos de machine learning permiten identificar patrones complejos en los datos de clientes que serían difíciles de descubrir mediante métodos tradicionales. Variables como la permanencia del cliente en la marca y el tiempo del cliente en la marca demostraron ser factores determinantes para predecir el riesgo de abandono. Esto sugiere que los clientes que han estado más tiempo con la marca, o aquellos que recientemente han realizado una compra, tienen menos probabilidades de abandonarla, mientras que aquellos con menor tiempo de permanencia o sin compras recientes están en mayor riesgo.

La capacidad de predecir el abandono con alta precisión tiene importantes implicaciones para la

estrategia de marketing y fidelización de las empresas de retail. Tradicionalmente, las estrategias de retención de clientes se han basado en datos históricos, como la recencia de la última compra, lo que resultaba en un enfoque reactivo. Sin embargo, con la implementación de este modelo predictivo, es posible anticiparse al comportamiento de los clientes y diseñar campañas de marketing personalizadas para aquellos que tienen mayor probabilidad de abandonar la marca. De esta forma, se pueden aplicar medidas preventivas, como ofertas personalizadas, promociones o programas de fidelización, antes de que los clientes dejen de interactuar con la empresa.

Otra conclusión importante es que, aunque todos los modelos utilizados mostraron un buen desempeño, la elección del modelo más adecuado dependerá de las necesidades específicas de la empresa. El modelo XGBoost demostró tener una ligera ventaja en cuanto a precisión y sensibilidad, pero también se consideraron otros factores, como el tiempo de entrenamiento y la complejidad computacional. En este sentido, el Random Forest, con una precisión del 85.86%, también podría ser una alternativa viable, especialmente si se busca un modelo más interpretable y menos costoso en términos computacionales. Las redes neuronales artificiales (ANN), aunque presentaron resultados sobresalientes, requirieron un tiempo de ejecución considerablemente mayor, lo que puede ser una limitante en entornos donde se necesita una predicción rápida y eficiente.

5.2 TRABAJOS FUTUROS

Para futuros trabajos relacionados con la predicción del riesgo de abandono de clientes en la industria del retail de moda, hay varias líneas de investigación y desarrollo que podrían explorar mejoras, optimizaciones y nuevas aplicaciones basadas en los resultados obtenidos en el proyecto actual. A continuación, se destacan algunas de las direcciones que podrían tomarse y los aspectos a considerar en cada una de ellas:

5.2.1. Mejora y expansión del modelo predictivo

Si bien el modelo actual basado en XGBoost ha mostrado un excelente desempeño, siempre es posible optimizar aún más el proceso de modelado y predicción. Algunos posibles caminos incluyen:

- Incorporación de nuevas variables: Se podrían incluir variables adicionales que no se utilizaron en este modelo, como el comportamiento en redes sociales, el historial de atención al cliente, reseñas en línea o incluso datos externos como tendencias económicas y de mercado. Estas variables podrían aportar nuevas perspectivas sobre los factores que influyen en la lealtad y el abandono del cliente.

- Ajuste de hiperparámetros: Aunque se han ajustado algunos hiperparámetros de los modelos, se podría realizar un ajuste más exhaustivo utilizando técnicas como la búsqueda en cuadrícula (Grid Search) o la optimización bayesiana, con el fin de mejorar la precisión y reducir el tiempo de ejecución de los modelos.

Experimentación con nuevos algoritmos: Además de los modelos utilizados, en trabajos futuros se podrían explorar otros algoritmos como las redes neuronales profundas (Deep Learning), redes neuronales recurrentes (RNN), o incluso algoritmos de ensamblado como LightGBM o CatBoost, que podrían mejorar aún más la capacidad predictiva del modelo, especialmente si se tiene acceso a mayores cantidades de datos.

- Modelos híbridos: Otra posible extensión sería combinar varios modelos en un enfoque híbrido, donde las predicciones de diferentes algoritmos se fusionen para obtener una predicción final más robusta. Los enfoques de ensamblaje como el *stacking* o *blending* podrían aplicarse para combinar las fortalezas de múltiples algoritmos.

5.2.2. Modelado dinámico y en tiempo real

Otro camino relevante sería crear modelos que puedan actualizarse de forma dinámica y hacer predicciones en tiempo real. Actualmente, el modelo se entrena de manera periódica (cada tres meses), pero los trabajos futuros podrían explorar:

- Actualización continua del modelo: Implementar un sistema de aprendizaje continuo que permita al modelo ajustarse en tiempo real conforme se incorporan nuevos datos. Esto requeriría diseñar flujos de trabajo que permitan la ingesta y procesamiento constante de datos en la producción.

- Predicciones en tiempo real: Integrar las predicciones en las interacciones en tiempo real con los clientes, permitiendo ofrecer incentivos de retención instantáneos (como descuentos u ofertas personalizadas) a aquellos que el modelo identifique como en riesgo de abandonar mientras navegan en el sitio web o interactúan con otros canales de venta.

5.2.3. Personalización de estrategias de retención

El modelo actual predice el riesgo de abandono, pero un posible trabajo futuro sería combinar esa información con sistemas de recomendación personalizados. Es decir, desarrollar modelos que no solo predigan el riesgo, sino que también sugieran la mejor estrategia de retención para cada cliente basado en sus preferencias y comportamiento.

- Sistemas de recomendación basados en el riesgo de abandono: Se podría investigar la implementación de un sistema de recomendación que ofrezca promociones, productos o servicios específicamente diseñados para retener a aquellos clientes con mayor riesgo de abandono. Esta personalización incrementaría la efectividad de las campañas de retención.

- Testeo A/B de estrategias: Los futuros trabajos podrían implementar pruebas A/B para evaluar qué tipos de incentivos o estrategias de retención son más efectivas en diferentes segmentos de clientes identificados como de alto riesgo por el modelo.

Bibliografía

[1] D. Popović, *Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm,* 2008.

[2] A. Pail, M. Deepshika, S. Mittal, S. K. Shetty, S. S. Hiremath y Y. E. Patil, *Customer Churn Prediction for Retail Business.,* 2017.

[3] A. Dingli, M. Vincent y S. F. Nicole, *Comparison of Deep Learning Algorithms to Predict Customer Churn within a Local Retail Industry.,* 2017.

[4] A. F. Echeverri Giraldo, *Modelo Predictivo de Churn de clientes para el negocio de Telecomunicaciones,* Medellín, 2019.

[5] J. D. Falla Arango, *Predicción de Abandono de Clientesen Telecomunicaciones Mediante Aprendizaje Automático,* Bogotá, 2021.

[6] E. A. Galvis Moncaleano, *Modelo de Churn para retención de clientes de Seguros Voluntarios,* Bogotá, 2023.

[7] J. Arango y D. Sánchez, «La predicción de abandono en la industria retail moda: un estudio de caso. Revista de Negocios,» p. 27, 2022.

[8] T. Hastie, R. Tibshirani y J. Friedman, «The elements of statistical learning., » Springer, 2009.

[9] S. Kumar y D. Chandrakala, «A survey on customer churn prediction using machine learning Techniques,» *International Journal of Computer Applications,* vol. 154, nº 10, 2016.

[10] T. Mitchell, Machine learning, New York: McGraw-Hill, 1997, p. 173.

[11] V. N. Vapnik, «The Nature of Statistical Learning Theory» *Springer,* 1995.

[12] C. Cortes y V. Vapnik, «Support-vector networks» *Machine Learning,* vol. 20, nº 3, pp. 273-297, 1995.

[13] D. Hosmer, S. Lemeshow y R. Sturdivant, Applied Logistic Regression, Wiley, 2013.

[14] L. Breiman, «Random forests» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[15] A. Liaw y M. Wiener, «Classification and Regression by random Forest., » *R News,* vol. 2, nº 3, pp. 18-22, 2002.

[16] J. Friedman, «Greedy function approximation: A gradient boosting machine., » *Annals of Statistics,* vol. 29, nº 5, pp. 1189-1232, 2001.

[17] A. Natekin y A. Knoll, «Gradient boosting machines, a tutorial» *Frontiers in Neurorobotics,* p. 7, 2013.

[18] S. Haykin, Neural Networks and Learning Machines, New York: 3r, 2009.

[19] I. Goodfellow, Y. Bengio y A. Courville, Deep Learning, Cambridge: MIT Press, 2016.

[20] C. Catal, «Performance Evaluation Metrics for Software» *Acta Polytechnica Hungarica,* vol. 9, nº 4, pp. 195-200, 2012.

[21] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «The KDD process for knowledge discovery in databases» *Communications of the ACM,* vol. 39, nº 11, pp. 27-34., 1996.

[22] K. J. Cios, W. Pedrycz y S. Swiniarski, «Data mining techniques: A practical approach» Springer, 2007.

[23] D. J. Hand, H. Mannila y P. Smyth, «Principles of data mining» MIT Press., 2001.
[24] I. H. Witten, E. Frank, M. Hall y C. Pal, Data mining: Practical machine learning tools and techniques., Morgan Kaufmann, 2016.
[25] M. A. Berry y G. S. Linoff, Mastering data mining: The art and science of transforming information into knowledge, Wiley., 2000.
[26] A. Vargas, «Multicolinealidad,» *Revista Colombiana de Estadística,* nº 2, 1980.