ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

MODELO PREDICTIVO DE CLIENTES DESERTORES. CASO DE ESTUDIO EN EL SECTOR DE SERVICIOS AUXILARES DEL SISTEMA FINANCIERO.

Profesor
Mario Salvador González

Autores
Arias Altamirano Diego Fernando
Cuaspud Torres Jackson Manuel

2024

RESUMEN

Una de las actividades más importantes en la gestión comercial de cualquier empresa es el manejo eficiente de su cartera de clientes. El éxito, tanto comercial como financiero, está estrechamente vinculado con la calidad y gestión de dicha cartera. A menudo, las empresas invierten considerables recursos financieros, materiales y humanos en captar nuevos clientes. Sin embargo, un aspecto clave que suele pasarse por alto es el análisis del comportamiento de los clientes actuales, incluyendo su nivel de fidelización, tasa de retención y, en última instancia, el índice de deserción.

Aunque a simple vista puede parecer una tarea sencilla, en el entorno actual de negocios, resulta complejo identificar con precisión a los clientes desertores y calcular su tasa de supervivencia o retención. Esto se debe a la gran cantidad de variables, tanto numéricas como categóricas, que influyen en diferentes grados sobre este tipo de clasificaciones. En este contexto, herramientas como el Machine Learning juegan un papel fundamental. A través de diversos algoritmos, estas tecnologías permiten procesar grandes volúmenes de datos de naturaleza diversa, facilitando el análisis detallado de la información clave para la toma de decisiones.

El objetivo de implementar estas herramientas es crear estrategias comerciales efectivas que reduzcan la deserción de clientes identificados como de alta probabilidad de fuga. Esto permitirá diseñar campañas específicas para retener a estos clientes, mejorando significativamente la tasa de retención y, en consecuencia, la supervivencia de la cartera. Al hacerlo, la empresa no solo optimiza su desempeño económico y financiero, sino que también fortalece su crecimiento sostenible y su posicionamiento competitivo en el mercado.

Palabras claves: Deserción Clientes, Supervivencia Clientes, Random Forest, Regresión Logística.

ABSTRACT

One of the most important activities in the commercial management of any company is the efficient handling of its customer portfolio. Success, both commercial and financial, is closely linked to the quality and management of that portfolio. Often, companies invest considerable financial, material, and human resources in acquiring new customers. However, a key aspect that is often overlooked is the analysis of current customer behavior, including their level of loyalty, retention rate, and ultimately, churn rate.

Although it may seem like a simple task at first glance, in today's business environment, identifying churned customers and accurately calculating their survival or retention rate can be quite complex. This is due to the vast number of variables—both numerical and categorical—that influence these classifications to varying degrees. In this context, tools like Machine Learning play a fundamental role. Through various algorithms, these technologies enable the processing of large volumes of diverse data, facilitating detailed analysis of the key information needed for decision-making.

The goal of implementing these tools is to create effective commercial strategies that reduce the churn of customers identified as high-risk. This allows for the design of targeted campaigns to retain these customers, significantly improving the retention rate and, consequently, the survival of the portfolio. By doing so, the company not only optimizes its economic and financial performance but also strengthens its sustainable growth and competitive positioning in the market.

Keywords: Customer Churn, Customer Survival, Random Forest, Logistic Regression.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCION

El avance tecnológico ha transformado radicalmente la forma en que se realizan muchas operaciones financieras cotidianas. Actividades que antes requerían desplazarse a una sucursal bancaria o a una institución especifica, como pagar servicios básicos o realizar depósitos, se simplificaron con la creación de puntos de pago, corresponsales no bancarios y puntos de venta. Estos canales no solo ofrecieron mayor cobertura geográfica, sino que también redujeron significativamente los tiempos de espera para los usuarios, facilitando el acceso a servicios financieros en zonas rurales y urbanas. La pandemia de COVID-19 acentuó la relevancia de estos servicios, ya que se convirtieron en una herramienta clave para minimizar los riesgos de contagio al evitar aglomeraciones en bancos y oficinas de servicios. A medida que la población optaba por alternativas electrónicas, estos puntos de pago adquirieron mayor protagonismo, consolidándose como una parte esencial del ecosistema financiero.

Sin embargo, una vez superada la etapa más crítica de la pandemia, el comportamiento del consumidor se adaptó, y el uso de estos puntos de pago electrónicos pasó a ser parte de la nueva normalidad. El crecimiento de la demanda atrajo la entrada de nuevas empresas al mercado, intensificando la competencia en un sector que ya enfrentaba retos considerables. Esta mayor competencia ha generado presión sobre las empresas que ofrecen servicios de intermediación financiera, pues ahora deben diferenciarse no solo en términos de cobertura, sino también en la calidad y estabilidad de sus plataformas. Con más alternativas disponibles, los puntos de venta pueden cambiar de proveedor con mayor facilidad, lo que aumenta el riesgo de deserción si no se satisfacen sus expectativas.

Para empresas dedicadas a la intermediación de pagos de servicios y transacciones electrónicas, mantener la lealtad de sus puntos de venta se ha convertido en un desafío estratégico. La deserción, de los establecimientos que actúan como puntos de venta, representa una amenaza significativa para la estabilidad del negocio, afectando no solo los ingresos por comisiones, sino también incrementando los costos operativos. Para enfrentar este reto, es crucial identificar las variables que influyen en la deserción, como la satisfacción del cliente, la capacidad de respuesta ante fallos en el servicio, las barreras al cambio, y la competencia en el mercado. Un análisis de estos factores puede ayudar a predecir patrones de comportamiento y diseñar estrategias que mejoren la retención de clientes, asegurando la sostenibilidad y el crecimiento en un sector tan dinámico y competitivo.

REVISIÓN DE LITERATURA

La temática de análisis de deserción de clientes es muy importante en las operaciones de las empresas sin importar el sector económico en el cual se desenvuelvan. (Gutiérrez González, 2020) menciona que el proceso de retención de un cliente es mucho más económico que captar uno nuevo, llegando incluso en algunos casos a ser “diez veces más barato” conservar que captar. Esta es la razón principal por la cual las estrategias de retención y fidelización deber ser aplicadas de forma constante.

Así también, el proceso de gestión de deserción de clientes – en ocasiones denominado fuga de clientes – es una actividad fundamental en todo departamento comercial; esta problemática se la aborda desde dos perspectivas: i) predecir los desertores potenciales y ii) desarrollar medidas preventivas y paliativas para evitar su ocurrencia. (Pinto Galindo, 2020). El reto que trae esto consigo, es que, en el contexto empresarial actual, esta información se encasilla en la definición de Big Data, por lo cual es necesario utilizar herramientas modernas de gestión de la información que permitan procesar y analizar la misma de forma eficiente.

La deserción (fuga) de clientes, conocida como Churn es una situación que se hace presente en prácticamente todas las actividades comerciales y empresariales. Las dinámicas del mercado, estrategias de marketing o insatisfacción de clientes, hacen que exista rotación de clientes. Sin embargo, esta situación debe ser observada con detenimiento, pues un indicador elevado de deserción de clientes – ya sea esporádico o recurrente - es una importante señal de alarma para la empresa. (Navas Ayala, 2023).

Para llevar a cabo de manera efectiva el análisis de deserción mencionado es fundamental utilizar herramientas como el aprendizaje automático, una técnica clave para identificar de forma automática patrones o tendencias dentro de un conjunto de datos. En los últimos años, esta técnica se ha vuelto esencial en casi todas las actividades que requieren extraer información valiosa de grandes volúmenes de datos. (Shalev-Shwartz & Ben-David, 2019)

En nuestra vida cotidiana, estamos constantemente interactuando con tecnología que utiliza aprendizaje automático: desde los sistemas de recomendación implementados en redes sociales o páginas de compras en línea, hasta la detección de datos biométricos en nuestros teléfonos móviles. Esta tecnología ha llegado a aplicarse de forma exitosa en múltiples sectores económicos como lo son el médico, logístico y transporte, publicidad, entre otros. (Maisueche Cuadrado, 2019)

Dada la complejidad y envergadura de las actividades que se deben desarrollar, es imposible para una persona programar de manera precisa todas las instrucciones necesarias para llevarlas a cabo. En su lugar, es esencial equipar a los sistemas computaciones con la capacidad de aprender por medio de la práctica y la experiencia, logrando de esta manera una capacidad de adaptación a nuevas circunstancias. (Shalev-Shwartz & Ben-David, 2019)

El objetivo que se busca alcanzar con la implementación de los diferentes modelos de predicción de clientes desertores es articular estrategias encaminadas a su retención y que, de ser efectivas, proporcionaran rendimientos económicos notables para la empresa. (Sierchuk, 2022) menciona si los algoritmos predictivos son aplicados y ejecutados de forma exitosa, pueden proporcionar a la empresa una serie de ventajas competitivas, pues estos brindan oportunidades de maximización de ingresos y reducción de costos.

(Ramirez, 2019) menciona los tres tipos de aprendizaje que hasta la actualidad se han desarrollado:

- El aprendizaje supervisado: se basa en un conjunto de datos que ha sido previamente etiquetados, lo que permite identificar patrones que luego pueden aplicarse para clasificar nuevos conjuntos de datos.
- El aprendizaje no supervisado: se utiliza cuando el conjunto de datos no posee etiquetas, por lo cual, con el fin de clasificarlos es necesario analizar las similitudes y diferencias para distinguirlos entre sí.
- El aprendizaje de refuerzo: se emplea cuando los datos del sistema no están etiquetados, pero, tras realizar diversas acciones y con el tiempo, el sistema recibe retroalimentación a través de actualizaciones.
- El aprendizaje profundo: Las redes de aprendizaje profundo funcionan detectando estructuras complejas en los datos que procesan. A través de modelos computacionales formados por múltiples capas de procesamiento, estas redes son capaces de generar varios niveles de abstracción para representar los datos.

Finalmente, el aprendizaje automático es una de las tecnologías más innovadoras que se pueden integrar en el departamento de análisis de información de la empresa. Mediante el uso de algoritmos, es posible analizar miles o incluso millones de datos de manera rápida y con mínima intervención humana. Esto permite identificar de manera oportuna factores como tendencias, indicadores y KPI en diversas áreas del negocio, transformando el proceso de análisis de una actividad meramente descriptiva a una predictiva.

En el presente trabajo, abordaremos la temática de la deserción de los clientes desertores desde dos enfoques: el enfoque estadístico tradicional y el machine learning; esto con el fin de realizar una comparación de su rendimiento y finalmente, seleccionar el que mejor se ajuste a los requerimientos específicas de la empresa objeto del estudio.

La regresión logística es un modelo de clasificación supervisada que se entrena con datos etiquetados para predecir la probabilidad de que ocurra un evento específico, como la probabilidad de cancelación de un servicio o la deserción de un cliente. Aunque es más sencillo que otros modelos complejos de aprendizaje automático, sigue siendo una herramienta valiosa y ampliamente utilizada en tareas de clasificación binaria, gracias a su clara interpretación y su eficacia en situaciones donde las relaciones entre variables son aproximadamente lineales. (Bahamonde Morales & Tapia Pisarro, 2022)

Por otro lado, es importante realizar una división de dataset con el fin de entrenarlo y probarlo, se la denomina Técnica de prueba dividida de entrenamiento (Train-Test Split). (Navas Ayala, 2023) indica que este método divide de forma aleatoria el dataset en dos partes, uno denominado de entrenamiento – train – y otro de prueba – test -. La primera partición, la de entrenamiento, por generalidad representa entre el 70% y 80% del dataset, el cual es utilizado para entrenar el modelo de machine learning. La segunda partición, tiene el restante 20% al 30% de los datos, y es utilizado para poner a prueba el algoritmo y verificar el nivel de ajuste del modelo.

El objetivo del Train-Test Split es garantizar que el modelo pueda generalizar bien a datos nuevos e invisibles, en lugar de simplemente aprender las peculiaridades del conjunto de datos original.

Dentro de los modelos de machine learning, el Random Forest tiene como característica principal el hecho que agrupa múltiples árboles de decisión para generar una predicción conjunta. Este método implica la creación de varios árboles a través de un re-muestreo de los datos originales, lo que permite que cada árbol sea entrenado con diferentes subconjuntos de datos. Además, también se selecciona aleatoriamente un subconjunto de las variables de entrada o explicativas para determinar cuáles serán utilizadas en cada árbol, contribuyendo así a la diversidad del conjunto de árboles y mejorando la robustez del modelo final (Beltrán & Barbona, 2022).

Como detalla (Javier, 2020) entre las ventajas del uso de un modelo de Random Forest tenemos:

- En el caso de clasificación, cada árbol emite un "voto" por una clase, y el modelo final asigna la clase que recibió más "votos". Para cada nueva observación, se recorre cada árbol y se elige la clase con la mayoría de los votos.
- En el caso de predicción, el modelo entrega un resultado que es el promedio de las salidas de todos los árboles.
- Este modelo es relativamente sencillo de entrenar en y logra un rendimiento igual de bueno que modelos más complejos.
- Tiene un rendimiento eficiente y se destaca como una de las técnicas más precisas cuando se trabaja con grandes sets de datos.
- El modelo es capaz de manejar un gran número de variables predictoras sin excluir ninguna, y también puede identificar cuáles son las más importantes, lo que lo hace útil para la reducción de dimensionalidad.
- Mantiene su precisión incluso cuando hay una gran cantidad de datos perdidos.

Sin embargo, también existen algunas desventajas como:

- Puede ser complicado interpretar los resultados cuando se presentan de manera gráfica.
- El modelo puede ajustarse en exceso a ciertos grupos de datos si hay ruido presente.
- Las predicciones no son continuas y el modelo no puede prever valores fuera del rango de datos utilizados para entrenarlo. Además, cuando se tienen predictores categóricos con diferentes números de niveles, los resultados pueden estar sesgados hacia aquellos niveles superiores.
- Hay poco control sobre el funcionamiento interno del modelo.

Una vez elaborado el modelo y ejecutadas las predicciones, en los dos subconjuntos de datos (entrenamiento y prueba), es necesario correr sobre el mismo una serie de evaluaciones o pruebas con el fin de determinar el nivel de ajuste y exactitud. (Orenes Casanova, 2022), menciona que para evaluar el rendimiento de un modelo de machine learning basado en la clasificación es necesario identificar con claridad el nivel de precisión o exactitud en las predicciones – clasificaciones – que han sido realizadas. Para este propósito, existen una serie de herramientas que, independientemente del modelo seleccionado, son de gran utilidad.

(Borja-Robalino, Monleón-Getino, & Rodellar, 2020) por su lado indican que las métricas de rendimiento son fundamentales en problemas de clasificación, ya que permiten evaluar y comparar la eficacia de diferentes algoritmos de Machine Learning y Deep Learning. Estas métricas no solo ayudan a determinar qué modelo es más preciso o eficiente, sino que también facilitan la selección del algoritmo más adecuado según el objetivo específico de la investigación. Al emplear métricas como la precisión, el recall, la F1 score y el área bajo la curva (AUC), los investigadores pueden discriminar entre algoritmos que, aunque puedan parecer similares en cuanto a su estructura, ofrecen resultados distintos en función del conjunto de datos y las metas del estudio. Así, la elección del algoritmo óptimo se basa en un análisis riguroso y cuantitativo que maximiza la relevancia y aplicabilidad de los resultados obtenidos.

La matriz de confusión es una de las herramientas más importantes y ampliamente utilizadas para evaluar el rendimiento de los modelos de clasificación, ya que permite realizar una comparación detallada entre las predicciones generadas por el algoritmo y los valores reales observados en el conjunto de prueba. En esencia, esta matriz organiza los resultados en cuatro categorías fundamentales: verdadero positivo (TP), falso positivo (FP), verdadero negativo (TN) y falso negativo (FN). Estas categorías indican si las predicciones coinciden o no con los valores reales. (Orenes Casanova, 2022).

Una vez realizada la Matriz de Confusión, podemos calcular con facilidad algunos indicadores que nos permitirán identificar de mejor manera el rendimiento del modelo, entre las cuales podemos mencionar los indicadores de Exactitud, Exhaustividad o Sensibilidad (también es conocido como Recall), Precisión (también conocido como Accuracy) y el F-Score.

El F-Score es una métrica crucial en la evaluación de modelos de clasificación, ya que combina de manera armoniosa la precisión (precision) y el recall en una única medida. A diferencia de la precisión (accuracy), que solo refleja la proporción de predicciones correctas sobre el total de predicciones, el F1-Score ofrece una visión más equilibrada del rendimiento, especialmente en escenarios donde las clases están desbalanceadas. (Borja-Robalino, Monleón-Getino, & Rodellar, 2020)

Otra herramienta de evaluación de los modelos es la Curva ROC (AUC – ROC), (Navas Ayala, 2023) indica que permite evaluar el desempeño del modelo de forma gráfica, mostrando la relación entre la Tasa de Verdaderos Positivos (TPR o *recall*) y la Tasa de Falsos Positivos (FPR) para cada posible umbral, generando una visualización bidimensional donde ambos ejes van de 0 a 1. De

tal manera que se podrá visualizar como el modelo responde en sus predicciones a medida que se ajusta el umbral de decisión.

Al analizar la evaluación de un modelo propuesto de Radom Forest se puede llegar a un nivel excesivo de sobreajuste por lo cual es necesario la inclusión de Hiperparametros como mencionan (Yang & Shami, 2020) tenemos:

- Función de medición: La función de medición evalúa la pureza de los nodos tras cada división y ayuda a seleccionar la mejor partición en cada etapa del entrenamiento del árbol. El índice de Gini mide la impureza de un nodo: un valor de 0 indica que el nodo es completamente puro. Por otro lado, la entropía evalúa la homogeneidad de los datos en un nodo de clasificación: un valor bajo o nulo de entropía sugiere que los datos son similares entre sí (Rivero, 2022).
- Número de árboles de decisión que hay que combinar: Se refiere a la cantidad de árboles individuales que se construyen y se combinan para formar el modelo (Yang & Shami, 2020).
- Número de características a considerar al buscar la mejor división: Este parámetro define cuántas de las características disponibles deben evaluarse en cada división de los nodos del árbol (Yang & Shami, 2020).
- Número mínimo de puntos de datos para dividir un nodo de decisión: Encuentra el mínimo de observaciones requeridas en un nodo para que pueda ser dividido. Un valor mayor de este parámetro reduce la flexibilidad del modelo (Serra, 2020).
- Número mínimo de puntos de datos para estar en un nodo hoja: Define la cantidad mínima de observaciones que debe tener un nodo para ser considerado como nodo hoja. Este parámetro tiene un efecto similar al número mínimo de observaciones para dividir un nodo (Serra, 2020).
- Profundidad máxima del árbol: Establece la profundidad máxima permitida en un árbol, entendiendo la profundidad máxima como el número de divisiones a lo largo de la rama más larga del árbol (Serra, 2020).

Para encontrar los ajustes óptimos de hiperparámetros, existen diversas técnicas, como expone (Alaminos, 2023). La búsqueda en cuadrícula que es una técnica que evalúa todas las combinaciones posibles de hiperparámetros dentro de un rango predefinido. En contraste, la búsqueda aleatoria selecciona un subconjunto aleatorio de combinaciones, lo que la hace más eficiente en términos de tiempo, al encontrar soluciones óptimas con menos iteraciones. Los métodos de optimización bayesiana utilizan información previa y actualizaciones basadas en datos para explorar el espacio de hiperparámetros de manera más

efectiva, adaptándose mejor a la estructura del espacio y mejorando la eficiencia en comparación con las búsquedas en cuadrícula y aleatoria.

El análisis de supervivencia como menciona (Paz, 2022) es una técnica estadística empleada para estudiar el tiempo que transcurre hasta que ocurre un evento específico, como la deserción de clientes. Debido a la presencia de datos censurados, ciertos estadísticos, como la media, pueden no ser apropiados ya que generan estimaciones sesgadas. En su lugar, se prefieren otros estimadores más robustos, como la mediana y la varianza. El objetivo principal del análisis es estimar los parámetros relacionados con la supervivencia y el riesgo, que determinan la distribución de los datos observados. Esto permite analizar no solo el tiempo de permanencia de los clientes, sino también los factores que influyen en estos procesos.

Una de las principales dificultades del análisis de supervivencia es trabajar con observaciones incompletas o censuradas. En estos casos, es necesario utilizar métodos adecuados para obtener estimaciones precisas de la supervivencia, como el estimador de Kaplan-Meier, que es ampliamente utilizado para manejar datos censurados. (Hernández, 2020) detalla que este estimador permite calcular la función de supervivencia a partir de los tiempos individuales de supervivencia de cada sujeto, lo que lo convierte en una herramienta precisa y efectiva. Al considerar los tiempos de supervivencia de manera individual, el estimador Kaplan-Meier ofrece una representación clara del comportamiento de la muestra.

Además, el estimador Kaplan-Meier es particularmente útil en análisis no paramétricos, permitiendo realizar comparaciones entre las curvas de supervivencia de dos o más grupos. Esto es esencial para entender cómo diferentes variables afectan la probabilidad de que ocurra un evento como la deserción o el egreso (Hernández, 2020).

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El presente trabajo tiene como objetivo analizar una serie de variables asociadas al perfil transaccional de los diferentes clientes de una empresa auxiliar de servicios financieros en el mercado ecuatoriano, con el fin de determinar si estos serán clientes desertores en el próximo período.

El dataset utilizado en este estudio proviene de datos históricos del nivel de ventas de los últimos doce meses de los clientes de la empresa en cuestión. Sobre estos datos, se han identificado variables relevantes que permiten analizar la correlación entre sus magnitudes, variaciones y tendencias, y la probabilidad de que estos clientes se conviertan en desertores en el futuro inmediato.

De esta manera, el objetivo del estudio es predecir, con una seguridad razonable, si un cliente será desertor o no, así también la tasa de supervivencia o retención en el periodo de doce meses posteriores a la fecha de evaluación. Para llevar a cabo esta predicción, se utilizarán algoritmos entrenados sobre un dataset, previamente analizado, depurado y estandarizado, que puedan ejecutarse sobre la nueva información que la empresa genere mes a mes y logren una detección adecuada de los potenciales clientes desertores y la expectativa de retención de la cartera de clientes en general.

Nos enfocaremos en los modelos de aprendizaje supervisado, pues como lo menciona (Arana, 2021) “estos tipos de modelos de aprendizaje supervisado se encuentran entre los modelos más utilizados y de mejor performance en la actualidad”, pues permiten resolver problemas de clasificación de una forma eficiente y relativamente sencilla de ejecutar, dado que los modelos son entrenados en función de etiquetas existentes en el dataset con el cual el modelo se entrena. Específicamente se ejecutarán dos modelos de predicción, por un lado una regresión logística y por otro random forest; así también, respecto del análisis de supervivencia de clientes, utilizaremos el modelo de Kaplan-Meier. El objetivo que se persigue es evaluar el desempeño de cada uno de ellos y determinar cuál se adapta de mejor manera a las especificidades de la organización.

La misión fundamental de este trabajo es proporcionar a la empresa una herramienta sólida que le permita estructurar y focalizar de forma eficiente los esfuerzos comerciales de retención de clientes, mediante una identificación adecuada de los mismos. Esto permitirá desarrollar estrategias personalizadas que satisfagan las necesidades y expectativas de los clientes, contribuyendo a un crecimiento sostenido y rentable de la operación de la empresa en el corto y largo plazo.

PLANTEAMIENTO DEL PROBLEMA

Desde una perspectiva amplia, la gestión comercial de clientes abarca dos actividades cruciales: la captación de nuevos clientes y la retención de los existentes. Muchas empresas concentran la mayoría de sus esfuerzos en la captación de nuevos clientes, una actividad vital para el crecimiento. Sin embargo, la retención de clientes a menudo queda desatendida, a pesar de ser fundamental para la consolidación y permanencia de las empresas en el mercado. La retención de clientes busca la fidelización, y un cliente fiel difícilmente abandonará la empresa, incluso frente a productos o servicios sustitutos más asequibles. Estos clientes valoran no solo la transacción económica, sino también características subjetivas que hacen a la empresa más atractiva que la competencia, como la reputación, garantía, atención al cliente, entre otros.

(Hernández, 2020) define la retención del cliente como el proceso en el cual un cliente insatisfecho se convierte en uno satisfecho y con el paso del tiempo en uno fidelizado respecto de los bienes o servicios que está adquiriendo. Sin embargo, otorga a la definición de insatisfecho una perspectiva más amplia y no necesariamente cargada de una connotación negativa, pues sigue una lógica simple, un cliente abandona – deja de consumir - la empresa por diferentes motivos que no han sido satisfechos a plenitud (precio, calidad, oportunidad, atención al cliente, estacionalidad, etc.).

Dada esta realidad, surge la necesidad de desarrollar estrategias para retener a los clientes. Aún más importante es identificar cuáles son los clientes más propensos a desertar o migrar hacia la competencia. ¿Cómo diferenciar a un cliente fidelizado, con baja probabilidad de deserción, de uno que potencialmente podría abandonar la empresa? ¿Cuáles son las variables o indicadores, tanto cualitativos como cuantitativos, que podrían ayudar a predecir esta tendencia? ¿Y bajo qué metodología se puede estructurar un proceso para identificar técnicamente a estos clientes?

Este es el desafío que enfrenta la empresa en estudio. Existe una gran oportunidad de mejorar la gestión de clientes —su retención y fidelización— mediante la identificación oportuna de clientes desertores, permitiendo desarrollar estrategias comerciales personalizadas. El primer paso es identificar a estos clientes de manera técnica, objetiva y estandarizada.

Ahora bien, este primer paso – la identificación de los posibles clientes desertores – trae consigo un reto particular, pues métodos tradicionales o empíricos que abordan esta problemática quedan obsoletos frente a la realidad

del mercado en el que se desenvuelve la empresa. La cantidad de variables que deben ser consideradas dentro del análisis, el volumen de información y la velocidad con la que esta se genera hace necesario implementar herramientas tecnológicas acordes a los desafíos actuales.

Por un lado, tenemos al Big Data, que constituye un cambio tecnológico significativo, puesto que los datos que se generan en las operaciones de las diferentes organizaciones lo hacen a gran velocidad, gran volumen, y con gran variedad. El Big Data por lo tanto, representa un profundo cambio de paradigma en la forma en que muchas – por no decir todas – las empresas realizan sus actividades. (Lee, 2019)

Otro concepto importante a considerar en este trabajo es el Machine Learning, que, por medio de diferentes algoritmos de aprendizaje, es capaz de generar modelos predictivos construidos a partir de patrones, tendencias y correlaciones entre diferentes variables presentes en el conjunto de datos a partir del cual se trabaja. (Shalev-Shwartz & Ben-David, 2019) mencionan que el machine Learning en la actualidad es una herramienta ampliamente utilizada para desarrollar aquellas actividades en las cuales se trabajen con grandes volúmenes de información.

En un mercado tan competitivo como el actual, la diferencia entre retener o perder a un cliente puede ser crucial, especialmente en el nicho de mercado donde opera la empresa en estudio. Los servicios auxiliares del sistema financiero obtienen su rentabilidad a través de márgenes repartidos en comisiones entre los diferentes actores del mercado. Esta característica específica añade un desafío adicional, ya que la rentabilidad se basa más en el volumen de transacciones y clientes que en el ARPU (ingreso promedio por usuario). Esta es una razón poderosa para abordar y resolver el problema de la deserción de clientes.

OBJETIVO GENERAL

Desarrollar un modelo predictivo utilizando técnicas de aprendizaje supervisado y herramientas de Big Data para identificar clientes con alta probabilidad de deserción, así como la tasa de supervivencia de los mismos, en una empresa ecuatoriana auxiliar de servicios financieros, con el fin de optimizar las estrategias de retención y fidelización de clientes.

OBJETIVOS ESPECÍFICOS

Analizar y depurar el dataset histórico de transacciones de clientes para identificar las variables relevantes que influyen en la deserción de clientes.

Implementar y entrenar un modelo de aprendizaje supervisado que utilice las variables identificadas para predecir la probabilidad de deserción de clientes, y su tasa de supervivencia, en períodos futuros.

Evaluar el modelo utilizando diversas métricas de rendimiento para garantizar su eficiencia, precisión y fiabilidad tanto en los datos de entrenamiento como en los de prueba, asegurando que, al ser implementado, proporcione un apoyo valioso para la gestión de la compañía.

Validar la efectividad del modelo predictivo mediante pruebas con datos recientes para asegurar su capacidad de identificar correctamente a los clientes con riesgo de deserción, así como su tasa de supervivencia – retención - .

Desarrollar una herramienta práctica para que la empresa pueda aplicar el modelo predictivo de manera continua sobre datos actualizados, permitiendo la detección temprana de clientes potencialmente desertores.

Proporcionar a la empresa una herramienta que facilite la generación de estrategias comerciales personalizadas basadas en los resultados del modelo, enfocadas en la retención y fidelización de los clientes identificados con alto riesgo de deserción.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Recolección de datos

Como se mencionó anteriormente, al ser una empresa dedicada a la intermediación de pagos de servicios y transacciones bancarias y electrónicas, la principal fuente de datos proviene de las diversas transacciones realizadas en sus plataformas propietarias o en el CRM. Estos datos pueden contener variables valiosas que ayudaran al desarrollo de un modelo de deserción de clientes.

En el primer paso, los datos se extraen de los diversos reportes disponibles lo cuales se convierten en tablas, que contienen información clave, como el número de transacciones, el valor de cada una, la categoría del cliente, fechas, ubicación, entre otros factores relevantes. Para este caso de estudio, se estructuró una base de datos en Power Pivot, integrando las distintas tablas, tal como se ilustra en el siguiente gráfico.

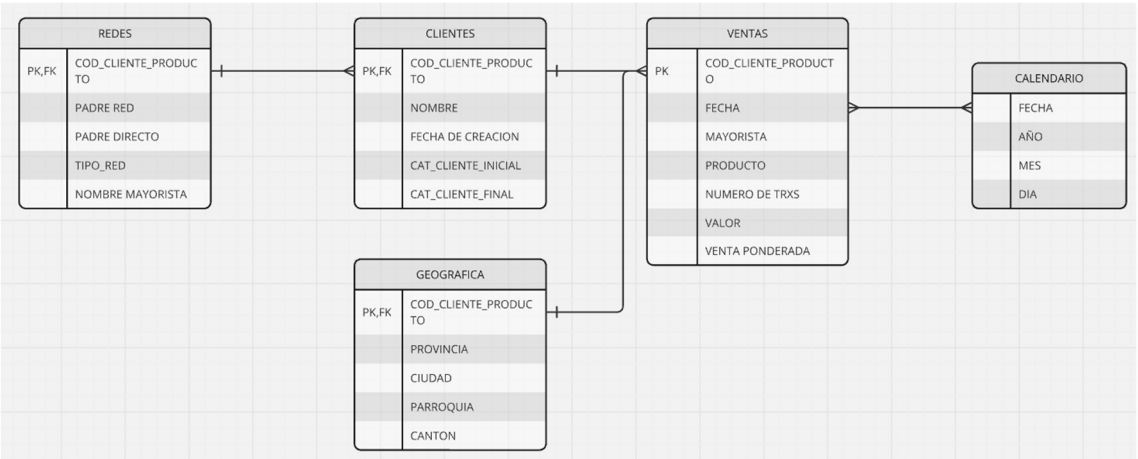

Tabla 1 – Tablas para construcción del Dataset

Para concluir, estructuramos nuestro dataset inicial realizando un mix de las distintas variables con conexión a las tablas, las cuales describen diversas características y comportamientos de los puntos de venta, entre ellas podríamos destacar las ventas representadas por cada mes, las categorías de los clientes, fecha de creación, región entre otros. Posteriormente, estas variables podrán ser transformadas para desarrollar un modelo más robusto.

Variables del Dataset Inicial	
Variable	**dtypes**
TIPO RED	int64
COD_CLIENTE_PRODUCTO	int64
Fecha Creación	datetime64[ns]
Region	object
JUN23	float64
JUL23	float64
AGO23	float64
SEP23	float64
OCT23	float64
NOV23	float64
DIC23	float64
ENE24	float64
FEB24	float64
MAR24	float64
ABR24	float64
MAY24	float64
JUN24	float64
Desertor	int64
CAT CLIENTE INICIAL	int64
CAT CLIENTE FINAL	int64
Antiguedad_Dias	int64

Tabla 2 – Variables del Dataset Inicial

Limpieza, pre-procesamiento y/o transformación de datos.

Una vez que se obtuvo el dataset que servirá de base para el presente trabajo, fue necesario ejecutar una serie de acondicionamientos adicionales con el fin de depurar, transformar y mejorar la información. El objetivo de este procedimiento es garantizar la calidad e integridad de la información contenida en el dataset y, consecuentemente, garantizar la exactitud de las predicciones que se realicen al ejecutar los algoritmos.

En primer lugar, se ejecutó un código que nos permitió transformar columnas con datos categóricos a datos numéricos (codificados), esto con el fin de facilitar la usabilidad de dichas variables al ejecutar los algoritmos.

Entre las variables categóricas modificadas podemos mencionar:

Nombre de la Variable	**TIPO RED**
Información que contiene	Clasificación de los puntos de venta (PDV) en función del canal con el cual fue creado: Directo – si la apertura fue sin intermediarios externos – o Mayorista – si la apertura se realizó por medio de agentes mayoristas.
Dato Original	**Dato Transformado**
Directo	1
Mayorista	0

Tabla 3 – Transformación de Datos – Variable: Tipo Red

Nombre de la Variable	**CAT CLIENTE INICIAL** **CAT CLIENTE FINAL**
Información que contiene	Clasificación de los puntos de venta (PDV) en función de su nivel de ventas mensuales al inicio y al final del periodo evaluado
Dato Original	**Dato Transformado**
C	1
B	2
A	3
AA	4
AAA	5
VIP	6

Tabla 4 – Transformación de Datos – Variable: Categoría Cliente

Nombre de la Variable	**REGION**
Información que contiene	Clasificación de los puntos de venta (PDV) en función de su ubicación geográfica dentro del territorio ecuatoriano
Dato Original	**Dato Transformado**
	Columnas Dumificadas
Costa	Region_Costa = {0,1}
Sierra	Region_Sierra = {0,1}
Oriente	Region_Oriente = {0,1}
Insular	Region_Insular = {0,1}

Tabla 5 – Transformación de Datos – Variable: Región

Posteriormente se procedió con la creación y transformación de las variables que contienen datos numéricos.

El dataset contiene doce columnas correspondientes a las ventas mensuales del último año de casa uno de los puntos de venta, en algunos casos, cuando el PDV no tenía ventas para dicho periodo el campo contenía un espacio en blanco, por lo que se procedió a rellenarlo con un cero, de esta forma se garantiza que los cálculos realizados a continuación no se distorsionen por valores faltantes.

Nombre de la Variable	**Antiguedad_Dias**
Información que contiene	Número de días transcurridos desde la fecha de creación del PDV hasta la fecha de corte.
Variable Fuente	Fecha Creación

Tabla 6 – Transformación de Datos – Variable: Antigüedad en días

Nombre de la Variable	**TASA_EVOLUCION_EXTREMOS**
Información que contiene	Porcentaje de variación de ventas entre el primer y último mes del periodo evaluado
Variable Fuente	JUN23 - MAY24

Tabla 7 – Transformación de Datos – Variable: Tasa Evolución Extremos

Nombre de la Variable	**Pendiente**
Información que contiene	Tendencia de crecimiento o decrecimiento esperada en función de las ventas mensualizadas
Variable Fuente	Doce columnas de ventas mensualizadas

Tabla 8 – Transformación de Datos – Variable: Pendiente

Nombre de la Variable	**TASA_DESERCION**
Información que contiene	Recuento de número de veces que un PDV desertó – dejó de transaccionar – en el periodo de doce meses evaluados.
Variable Fuente	Doce columnas de ventas mensualizadas

Tabla 9 – Transformación de Datos – Variable: Tasa Deserción

Nombre de la Variable	**TASA_EVOLUCION_MENSUAL_PROMEDIO**
Información que contiene	Promedio de la variación mensualizada de ventas en el periodo de doce meses evaluado.
Variable Fuente	Doce columnas de ventas mensualizadas

Tabla 10 – Transformación de Datos – Variable: Tasa Evolución Mensual

Nombre de la Variable	PROB_DESERCION
Información que contiene	Promedio de la probabilidad mensualizadas de actividad (100%) o deserción (0%) de cada PDV. En este caso se asigna como mes activo a partir de 0.01 ctv de ventas.
Variable Fuente	Doce columnas de ventas mensualizadas

Tabla 11 – Transformación de Datos – Variable: Probabilidad de Deserción

Nombre de la Variable	TASA_PERD_MAX
Información que contiene	Porcentaje de variación de ventas entre el mayor y menor mes de ventas del periodo evaluado
Variable Fuente	Mínimo y máximo de columnas de ventas mensualizadas

Tabla 12 – Transformación de Datos – Variable: Tasa Máxima de Pérdida

Una vez ejecutado estos cambios, obtenemos el dataset definitivo con el que trabajaremos los algoritmos. Finalmente, se verifica que no existan valores nulos o en blancos.

Porcentaje de Datos Perdidos

Variable	%	Variable	%
TIPO RED	0%	TASA_DESERCION	0%
COD_CLIENTE_PRODUCTO	0%	TASA_EVOLUCION_MENSUAL_PROM	0%
Desertor	0%	PROB_DESERCION	0%
CAT CLIENTE INICIAL	0%	TASA_PERD_MAX	0%
CAT CLIENTE FINAL	0%	Region_COSTA	0%
Antiguedad_Dias	0%	Region_INSULAR	0%
TASA_EVOLUCION_EXTREMOS	0%	Region_ORIENTE	0%
Pendiente	0%	Region_SIERRA	0%

Tabla 13 – Análisis de Datos Perdidos

Finalmente, se presenta a continuación el detalle del tipo de dato que contiene cada una de las columnas del dataset:

Tipo de Datos de las Variables del Dataset

Variable	dtypes	Variable	dtypes
TIPO RED	int32	TASA_DESERCION	float64
COD_CLIENTE_PRODUCTO	int64	TASA_EVOLUCION_MENSUAL_PROM	float64
Desertor	int64	PROB_DESERCION	float64
CAT CLIENTE INICIAL	int32	TASA_PERD_MAX	float64
CAT CLIENTE FINAL	int32	Region_COSTA	int32
Antiguedad_Dias	int32	Region_INSULAR	int32
TASA_EVOLUCION_EXTREMOS	float64	Region_ORIENTE	int32
Pendiente	float64	Region_SIERRA	int32

Tabla 14 – Tipo de Datos de las Variables del Dataset

Identificación y descripción de variables

Una vez realizado el proceso de exploración y transformación de las variables, obtuvimos el dataset final con el cual se procederá a correr los diferentes modelos.

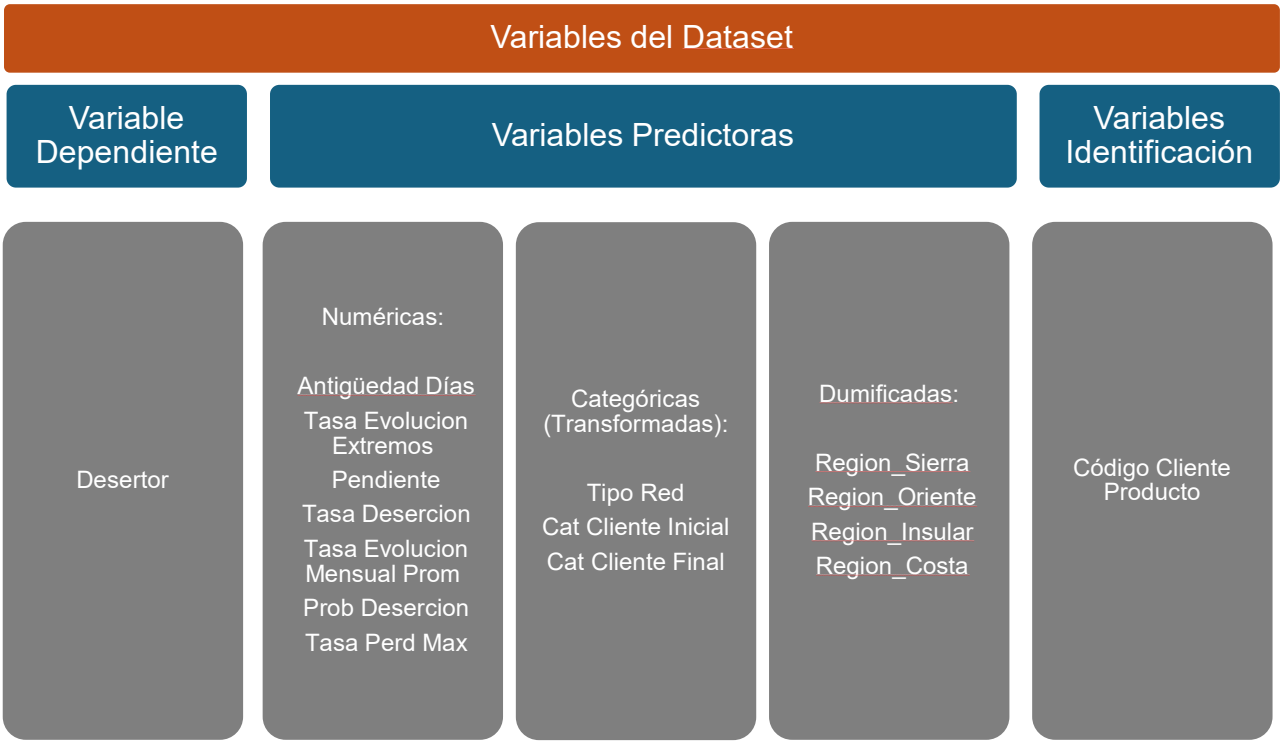

Tabla 15 – Clasificación de las Variables del Dataset

A continuación, se describen las variables que confirman el Dataset final:

Variable	Tipo	Naturaleza	Descripción	Fuente de datos
COD CLIENTE PRODUCTO	int64	Numérica	Identificador único correspondiente a cada Punto de Venta.	Sistema transaccional
Desertor	int64	Categórica (Transformada)	Indica con un 1 si el cliente es un desertor y con un 0 si no lo es.	Sistema transaccional
TIPO RED	Int32	Categórica (Transformada)	Segmenta los Puntos de Venta: 1 para los creados directamente y 0 para los creados a través de un mayorista intermediario.	Sistema transaccional
CAT CLIENTE INICIAL	Int32	Categórica (Transformada)	Segmentación de los Puntos de Venta al inicio del periodo de estudio según sus ventas: • C:1 • B:2 • A:3 • AA:4 • AAA:5 • VIP:6	Sistema transaccional
CAT CLIENTE FINAL	Int32	Categórica (Transformada)	Segmentación de los Puntos de Venta al final del periodo de estudio según sus ventas: Segmentación de los Puntos de Venta al inicio del periodo de estudio según sus ventas: • C:1 • B:2 • A:3 • AA:4 • AAA:5 • VIP:6.	Sistema transaccional
Region_ COSTA	Int32	Categórica (Dumificada)	Dummy que indica con un 1 si el Punto de Venta está en la Región Costa y con un 0 si no.	Sistema transaccional
Region_ INSULAR	Int32	Categórica (Dumificada)	Dummy que indica con un 1 si el Punto de Venta está en la	Sistema transaccional

			Región Insular y con un 0 si no.	
Region_ ORIENTE	Int32	Categórica (Dumificada)	Dummy que indica con un 1 si el Punto de Venta está en la Región Oriente y con un 0 si no.	Sistema transaccional
Region_ SIERRA	Int32	Categórica (Dumificada)	Dummy que indica con un 1 si el Punto de Venta está en la Región Sierra y con un 0 si no.	Sistema transaccional
Antigüedad_ Días	Int32	Numérica	Muestra el número de días transcurridos desde la apertura del Punto de Venta hasta la fecha de corte del estudio.	Sistema transaccional
TASA_ EVOLUCION_ EXTREMOS	float64	Continua	Describe la variación porcentual entre el primer y el último mes del periodo de estudio.	Sistema transaccional
Pendiente	float64	Continua	Indica la tendencia de crecimiento o decrecimiento de las ventas de un Punto de Venta.	Sistema transaccional
TASA_ DESERCION	float64	Continua	Tasa que describe las veces que un Punto de Venta dejó de transaccionar durante el periodo analizado.	Sistema transaccional
TASA_ EVOLUCION_ MENSUAL_ PROM	float64	Continua	Tasa que describe el promedio de variación de las ventas de forma mensual durante el periodo de estudio.	Sistema transaccional
PROB_ DESERCION	float64	Continua	Valor promedio de la probabilidad de actividad de forma mensual.	Sistema transaccional
TASA_ PERD_MAX	float64	Continua	Muestra el porcentaje de variación entre el mes de mayores ventas y el mes de menores ventas.	Sistema transaccional

Tabla 16 – Variables de Dataset Final

Visualización de variables

Previo a ejecutar los algoritmos de clasificación es importante explorarlos a fin de entender claramente la forma en que estos están distribuidos en el dataset. La mejor manera de realizarlos es por medio de la representación gráfica de las diferentes variables.

En primer lugar, analizamos la variable dependiente (predictora); de un total de 44.378 observaciones, dos tercios de los mismos son clientes que se mantienen – mantendrán – activos, mientras el tercio restante son clientes desertores. A nuestro criterio, la distribución de la variable dependiente es lo suficientemente razonable para evitar problemas de clases desbalanceadas.

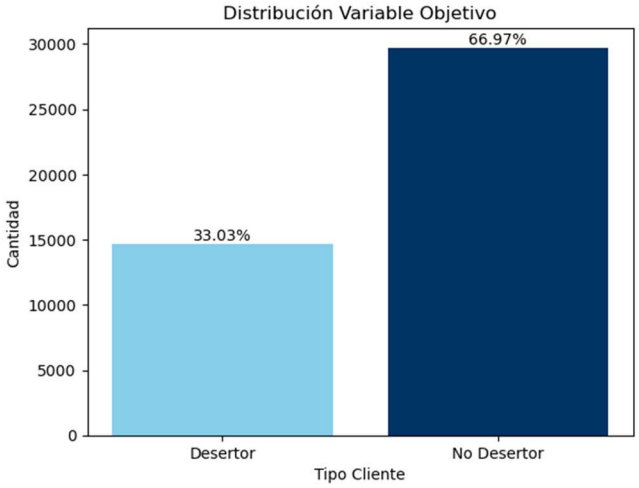

Figura 1 – Balanceo Variable Predictora

Así también se verificó la distribución de los puntos de venta en función de la categorización del nivel de ventas al inicio y final del periodo evaluado; se pudo identificar que la distribución entre las distintas categorías (de menor a mayor: C, B, A, AA, AAA, VIP) guarda las mismas proporciones.

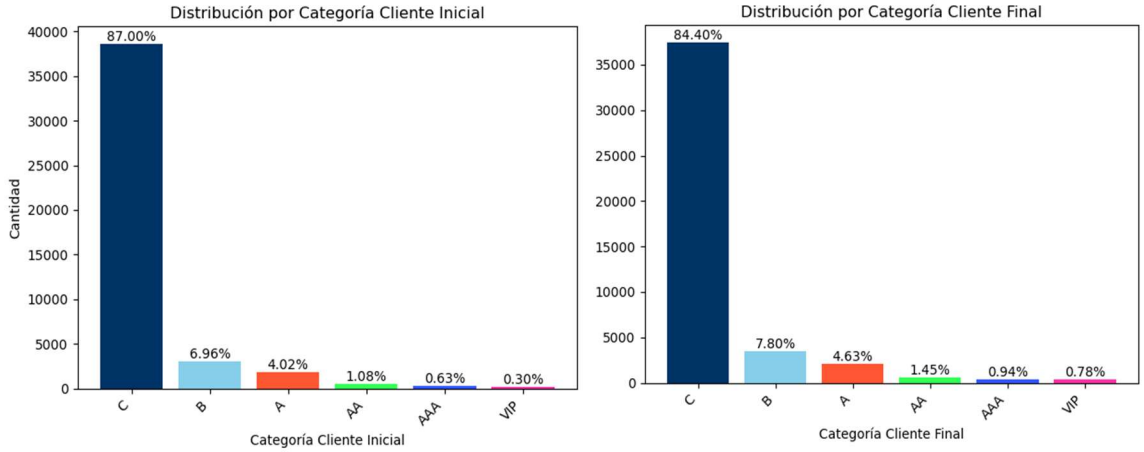

Figura 2 – Análisis Distribución Variables – Clientes por Categoría

Resulta interesante analizarlo desde otra perspectiva, esta vez, visualizando la distribución de las ventas en función de la categoría del cliente al inicio y final del periodo evaluado. En este caso, una categoría que tiene un número pequeño de clientes como es el VIP, que representa menos del 1% del total de clientes, logra participar fuertemente entre el 9.51% y 24.82% de las ventas, además el incremento de la participación se explica por la consolidación y fidelización de esta categoría en especial, motivado principalmente a la estrategia comercial hacia estos clientes por la importancia que poseen.

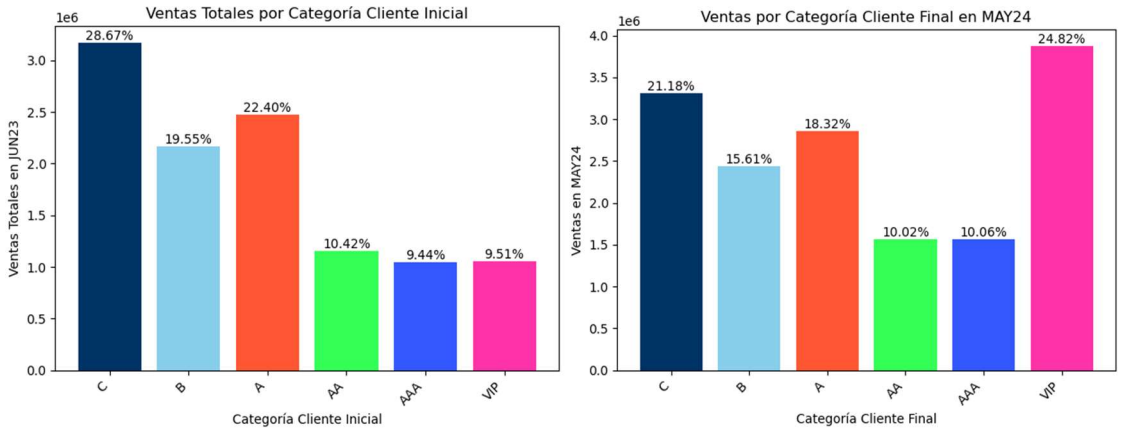

Figura 3 – Análisis Distribución Variables – Ventas por Categoría Clientes

Por otro lado, se analizó la distribución de los puntos de ventas en función de dos categorías: el tipo de red y su ubicación geográfica. Por el tipo de red tenemos una preponderancia de la Red Directa (aquellos PDV creados directamente por la empresa) mientras que la red Mayorista (PDV creados por un intermediario externo a la empresa) participa en menor medida.

Desde la perspectiva de la ubicación geográfica, la mayor cantidad de puntos de venta se concentran en la región sierra, seguido de la región costa. Esta distribución es adecuada si se tiene presente que: i) la empresa inició sus operaciones en la sierra y que recientemente se encuentra enfocada en extenderse hacia las demás provincias y ii) por distribución de población en términos generales, estas dos regiones son las más pobladas.

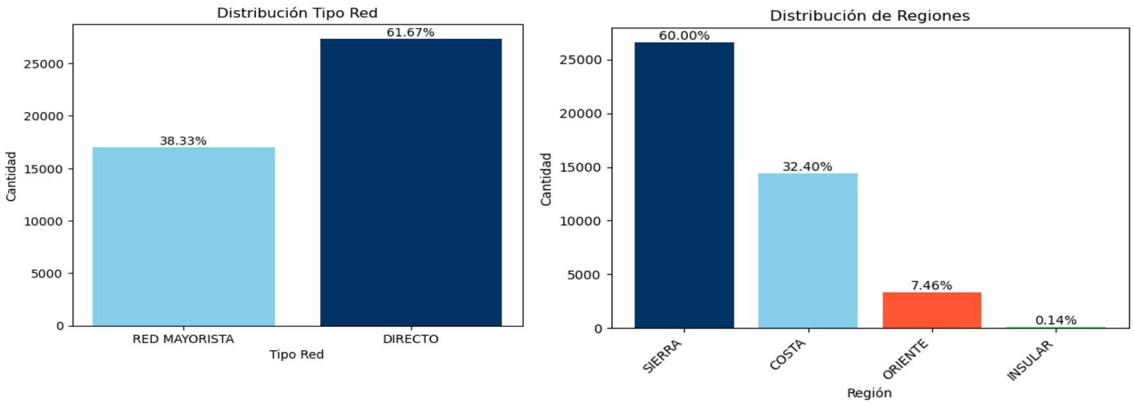

Figura 4 – Análisis Distribución Variables – Clientes por Región y Tipo Red

Por otro lado se realizó un análisis de la relación que tiene la variable dependiente (Desertor) con ciertas variables categóricas de nuestro dataset; en un primer acercamiento, se exploró la distribución que presenta con las variables de categoría de cliente.

Como se puede observar a continuación, la gran mayoría de deserción se encuentra en la categoría C, que corresponde a la categoría más numerosa de clientes. Así también, se puede evidenciar que en esta categoría se mantienen las proporciones de clasificación de clases en un 67% y 33% entre no desertor y desertor. Para el resto de categorías de clientes, la presencia de desertores es bastante baja.

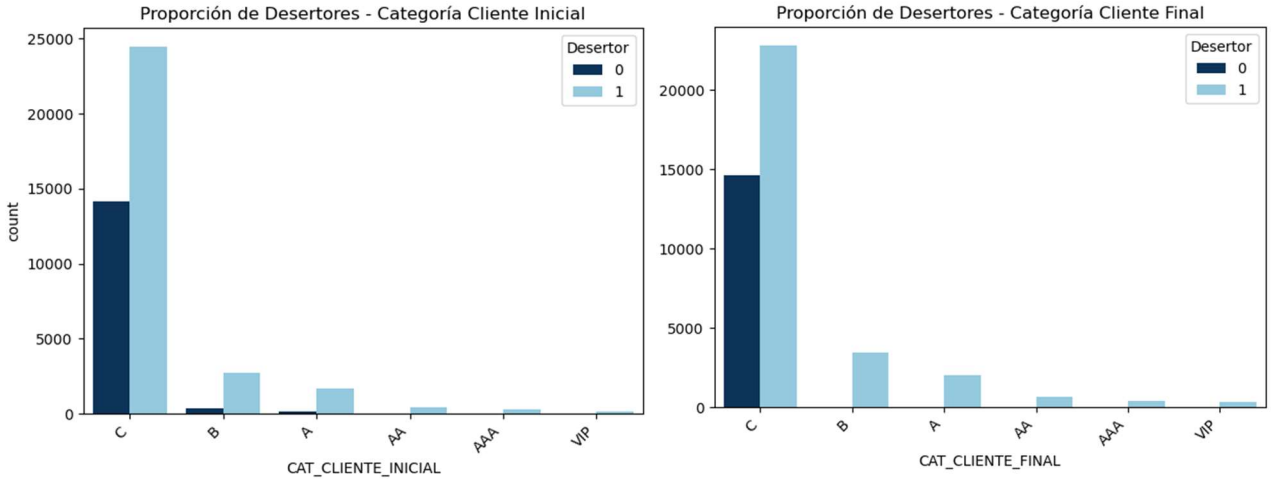

Figura 5 – Análisis Proporción Deserción por Categoría Cliente

También se analizó la relación de la variable “Desertor”, en relación a otras dos categorías (Región y Tipo Red), en la cual podemos identificar que la distribución de clases es bastante homogénea en la todas las categorías, descartando así un posible desbalanceo que pudiese afectar negativamente los modelos que serán implementados en el presente trabajo.

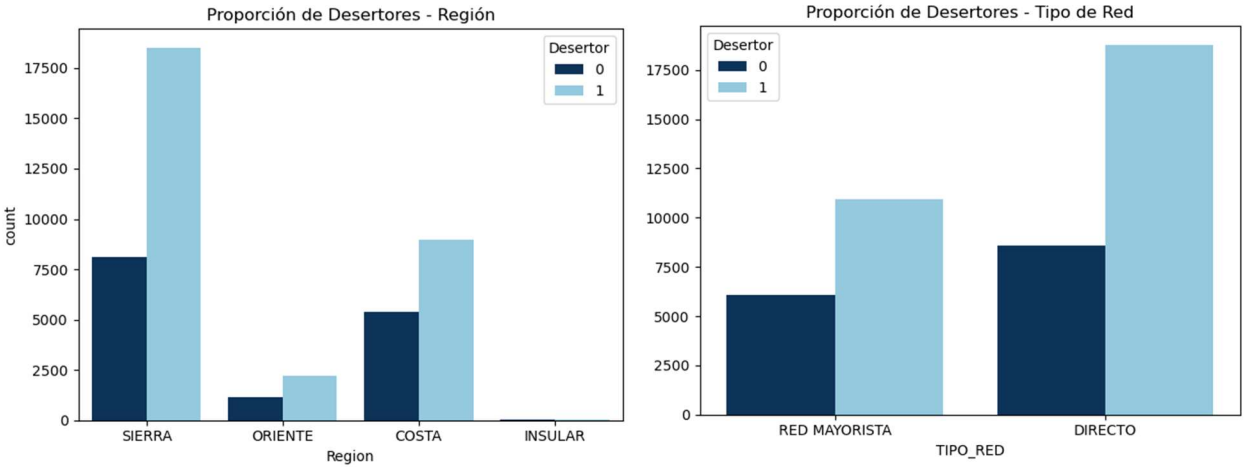

Figura 6 – Análisis Proporción Deserción por Región y Tipo Red

Finalmente, se presenta la matriz de correlación entre las dieciséis variables que conforman el datase depurado. Importante mencionar que nuestra variable dependiente, al ser binaria, no necesariamente presentará una relación lineal fuerte con otras variables.

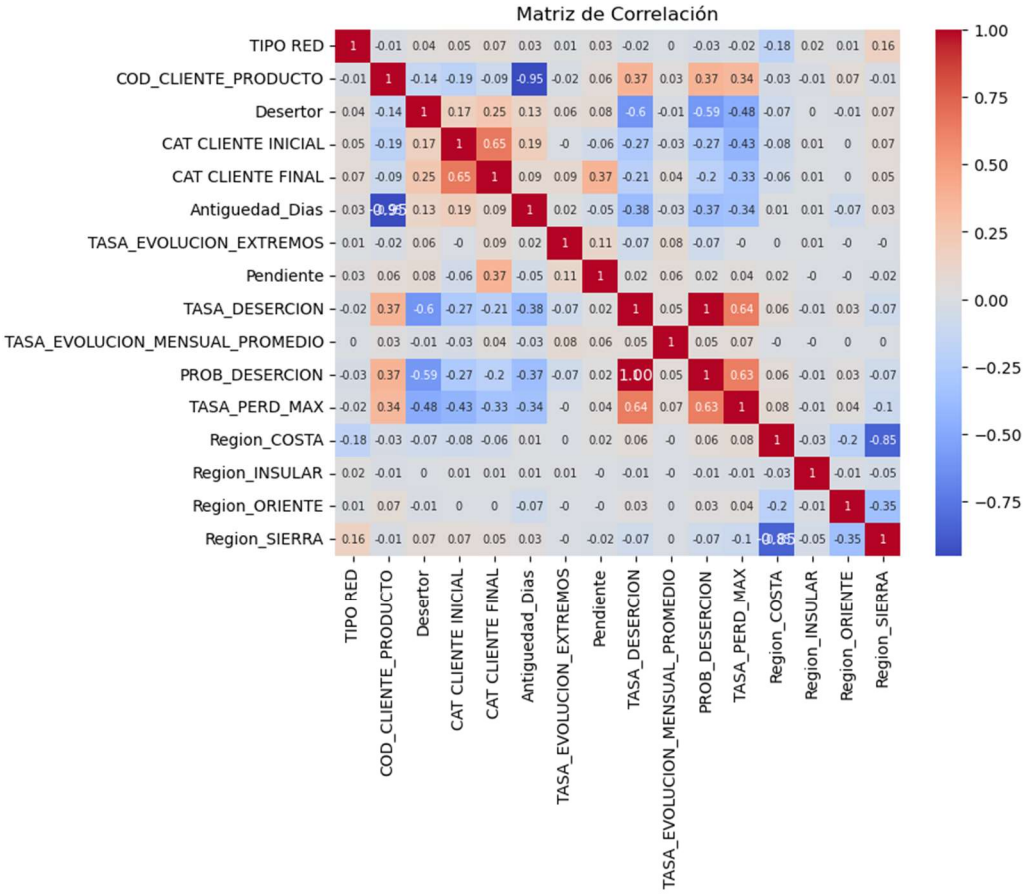

Figura 7 – Matriz de Correlación

Para una mejor interpretación de la variable dependiente con las variables independientes, a continuación, se presenta las diferentes correlaciones que fueron calculadas:

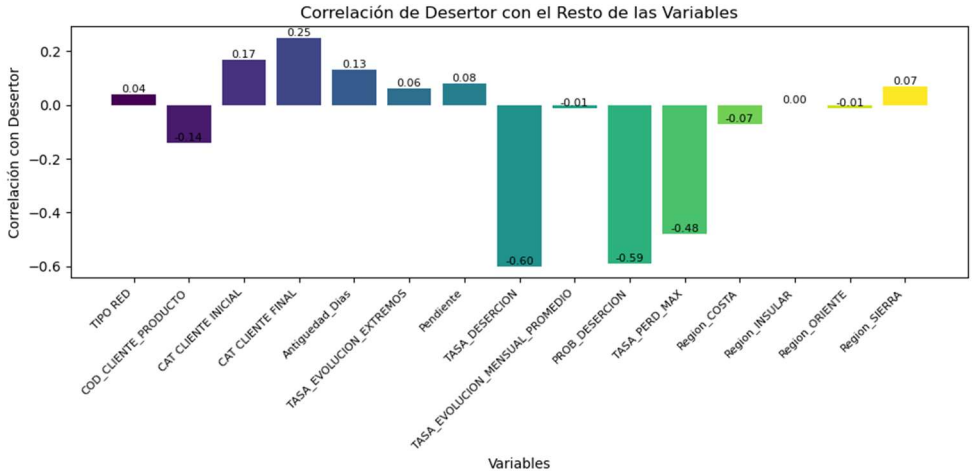

Figura 8 – Correlación de la Variable Predictora con el Dataset

Selección de modelo estadístico

El presente trabajo está enfocado en comparar e identificar un algoritmo de predicción de posibles clientes desertores que mejor se ajuste a las condiciones de la empresa seleccionada. Así también, como complemento, utilizaremos un algoritmo de análisis de supervivencia con el fin de no solo identificar los clientes desertores, si no también, realizar una estimación del tiempo en que este evento sucederá. Para la primera parte – identificación de clientes desertores – se utilizará dos enfoques: regresión logística y random forest; para la segunda parte analizaremos la tasa de supervivencia por medio de Kaplan-Meier.

En lo que respecta la regresión logística "se basa como método estadístico utilizado en la estimación de la probabilidad de ocurrencia en un evento específico. Este modelo analiza la relación entre las características o variables predictoras y calcula la probabilidad de obtener algún resultado particular" (Navas Ayala, 2023).

En nuestro caso particular, la variable predictora únicamente posee dos estados: cliente activo o desertor, por lo cual, al ser una variable binaria, utilizaremos específicamente el modelo de regresión logística binaria.

La función logística también conocida como Sigmoida generalmente se denota por σ(x) o sig(x) y está dada por:

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

Esta es la función matemática que genera la conocida "curva en forma de S", cuyo valor siempre se encuentra entre 0 y 1. Por esta razón, se utiliza para resolver problemas de clasificación binaria, es decir, aquellos en los que solo hay dos posibles categorías.

Además, se establece un punto de corte, comúnmente fijado en 0.5. De esta manera, si el valor predicho por la curva logística supera este umbral (p > 0.5), la nueva observación se clasifica como y=1; de lo contrario, se clasifica como y=0.

Tomando en cuenta la función logit, podemos definir a la ecuación logística como:

$$logit(p) = log(odds) = logit(Y) = \ln(\frac{p}{1-p}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots . \hat{\beta}_n X_n + \underbrace{\epsilon}_{error}$$

Por otro lado, tenemos a Random Forest, “un algoritmo de aprendizaje automático que se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y la combinación de sus predicciones para obtener un resultado final. Este enfoque permite reducir el sobreajuste y mejorar la capacidad de generalización del modelo, al tiempo que mantiene un buen rendimiento predictivo”. (Bahamonde Morales & Tapia Pisarro, 2022)

Este tipo de algoritmos opera bajo configuraciones o definiciones denominadas “hiperparámetros”, mismos que deben ser definidos de forma previa al entrenamiento de los datos, entre los hiperparámetros podemos mencionar tamaño del nodo, cantidad de árboles y cantidad de características muestreadas.

De forma matemática, el modelo de random forest, utiliza una técnica de bootstrap o bagging, trabajando sobre un conjunto de entrenamiento $X = x_1, \dots, x_n$ con una serie de respuestas $Y = y_1, \dots, y_n$, donde:

$$b = 1, \dots, B$$

Dicha muestra, al ser reemplazada en los datos de entrenamiento X, Y pasan a denominarse $X_b Y_b$ y se procede a entrenar el árbol de clasificación o regresión $f_b \, en \, X_b Y_b$.

Finalmente, las predicciones sobre los datos no vistos son obtenidos mediante el voto de pluralidad en el caso de árboles de clasificación, o mediante el promedio las predicciones de todos los árboles de regresión individuales.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Para concluir, examinaremos la tasa de supervivencia, que implica el análisis de datos mediante métodos adecuados para evitar sesgos en las estimaciones, dado que estadísticos como la media pueden resultar inexactos.

El estimador Kaplan-Meier es uno de los más utilizados para calcular la función de supervivencia, ya que permite manejar datos censurados y generar estimaciones precisas a partir de los tiempos de supervivencia individuales. Este método no paramétrico también facilita la comparación de curvas de supervivencia entre diferentes grupos, convirtiéndose en una herramienta eficaz para analizar la tasa de supervivencia y estimar parámetros clave.

El método Kaplan-Meier estima la probabilidad de supervivencia para cada tiempo en que ocurre un evento, proporcionando una función de supervivencia. El intervalo entre eventos se define desde el tiempo anterior al actual. Como menciona (Sánchez, 2021) en cada intervalo, se cuenta cuántos individuos sobreviven hasta ese momento y cuántos experimentan el evento. La probabilidad de supervivencia se calcula como la probabilidad de que el evento ocurra después de un tiempo determinado.

La probabilidad de supervivencia en un tiempo t_j está vinculada a la probabilidad de haber sobrevivido hasta el tiempo anterior t_{j-1}, y depende del número de individuos que aún no han experimentado el evento (r_j) y de los que lo han hecho (d_j). Esta probabilidad solo se calcula cuando ocurre el evento y se mantiene igual en los intervalos donde no hay eventos. La fórmula que describe esta probabilidad es

$$P = \frac{r_j - d_j}{r_j}$$

Si se lo realiza sucesivamente para todos los tiempos se logra una estimación de la curva de supervivencia.

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{r_j - d_j}{r_j}$$

Luego se aplicará la prueba log-rank que examina la igualdad entre las funciones de supervivencia de distintos grupos comparando el número de eventos observados con los esperados. La prueba se basa en el análisis del comportamiento global de las curvas de supervivencia a lo largo del tiempo.

RESULTADOS Y PROPUESTA DE SOLUCIÓN AL PROBLEMA IDENTIFICADO

Análisis de Deserción: Regresión Logística vs Random Forest

Con el objetivo de identificar con un grado de certeza razonable a los clientes potenciales que podrían desertar, se han seleccionado y evaluado dos modelos predictivos: la regresión logística y el algoritmo de Random Forest. Ambos enfoques permiten abordar el problema desde perspectivas complementarias: la regresión logística, como un modelo estadístico tradicional, facilita la interpretación de las probabilidades y relaciones lineales entre las variables; mientras que el Random Forest, como modelo basado en árboles de decisión, destaca por su capacidad para manejar interacciones no lineales y su resistencia al sobreajuste.

El propósito fundamental de este análisis es comprender las fortalezas y debilidades de cada uno de estos modelos, evaluando su desempeño en términos de precisión, sensibilidad, especificidad y capacidad de generalización. Al hacerlo, se busca seleccionar el modelo que mejor se ajuste a las características únicas del conjunto de datos disponible, el cual refleja las particularidades del negocio en estudio. Esta evaluación no solo permitirá una predicción más precisa de los desertores, sino que también proporcionará una base sólida para diseñar estrategias comerciales de retención más efectivas.

Análisis de modelo estadístico

Como primero paso analizaremos el enfoque de un modelo de regresión logística para el cual nos da como resultado el siguiente modelo, con las variables de mayor importancia que ayudará a predecir la deserción de los Puntos de Venta:

$$\begin{aligned} logit(P) = {} & \beta_0 + \beta_1 TIPO\ RED + \beta_2 CAT\ CLIENTE_{INICIAL} + \beta_3 CAT\ CLIENTE\ FINAL \\ & + \beta_4 Antiguedad\ Dias + \beta_5 Pendiente + \beta_6 TASA\ DESERCION \\ & + \beta_7 TASA\ EVOLUCION\ MENSUAL\ PROMEDIO \\ & + \beta_8 PROB\ DESERCION + \beta_9 TASA\ PERD\ MAX + \beta_{10} Region\_COSTA \end{aligned}$$

Logit Regression Results

Dep. Variable:	Desertor	No. Observations:	35502
Model:	Logit	Df Residuals:	35491
Method:	MLE	Df Model:	10
Date:	Tue, 24 Sep 2024	Pseudo R-squ.:	0.5583
Time:	14:21:55	Log-Likelihood:	-9926.9
converged:	True	LL-Null:	-22475.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.4156	0.465	13.796	0.000	5.504	7.327
TIPO_RED	0.1920	0.037	5.214	0.000	0.120	0.264
CAT_CLIENTE_INICIAL	1.5103	0.102	14.756	0.000	1.310	1.711
CAT_CLIENTE_FINAL	2.7568	0.256	10.784	0.000	2.256	3.258
Antiguedad_Dias	-0.0008	2.79e-05	-29.818	0.000	-0.001	-0.001
Pendiente	0.0334	0.001	25.963	0.000	0.031	0.036
TASA_DESERCION	-65.0812	1.230	-52.911	0.000	-67.492	-62.670
TASA_EVOLUCION_MENSUAL_PROMEDIO	0.1880	0.019	10.126	0.000	0.152	0.224
PROB_DESERCION	58.8037	1.148	51.208	0.000	56.553	61.054
TASA_PERD_MAX	-7.4178	0.388	-19.107	0.000	-8.179	-6.657
Region_COSTA	-0.0730	0.038	-1.925	0.054	-0.147	0.001

Tabla 17 – Resultados Regresión Logística

Para un mejor análisis se realizó la conversión de los coeficientes a Odds Ratio para una mejor interpretación.

	Odds Ratio
const	6.112936e+02
TIPO_RED	1.211622e+00
CAT_CLIENTE_INICIAL	4.528020e+00
CAT_CLIENTE_FINAL	1.574919e+01
Antiguedad_Dias	9.991674e-01
Pendiente	1.033949e+00
TASA_DESERCION	5.439700e-29
TASA_EVOLUCION_MENSUAL_PROMEDIO	1.206835e+00
PROB_DESERCION	3.452518e+25
TASA_PERD_MAX	6.004607e-04
Region_COSTA	9.295996e-01

Tabla 18 – Odds Ratios

El modelo de regresión logística demuestra que varias variables tienen un impacto significativo en la probabilidad de deserción de los clientes. El intercepto del modelo indica que, en ausencia de otras variables, las probabilidades de que un cliente deserte son considerablemente altas, con una razón de 611 veces más de probabilidad.

El TIPO_RED cuando el resto de las variables se mantienen constantes, un aumento en esta variable incrementa las probabilidades de deserción 1.21 veces. Y al ser una variable dicotómica se entiende que los puntos de venta de la Red Directa tienen mayores probabilidades de deserción.

Las variables CAT_CLIENTE_INICIAL y CAT_CLIENTE_FINAL cuando el resto de las variables se mantienen constantes, tienen un impacto mucho mayor: un aumento en la categoría del cliente inicial incrementa las probabilidades de deserción 4.53 veces, mientras que en la categoría final este incremento alcanza las 15.75 veces. Lo cual nos dice que conforme aumente la categoría del cliente puede aumentar la probabilidad de deserción.

La Antiguedad_Dias cuando el resto de las variables se mantienen constantes, reduce la probabilidad de deserción ya que cada día adicional reduce las probabilidades de deserción en 0.9992 veces. Con respecto a la Pendiente cuando el resto de las variables se mantienen constantes, aumenta en menor medida la probabilidad de deserción en 1.0339 veces.

La TASA_EVOLUCION_MENSUAL_PROMEDIO cuando el resto de las variables se mantienen constantes, influye, incrementando las probabilidades de deserción 1.21 veces por cada unidad de aumento. La TASA_DESERCION cuando el resto de las variables se mantienen constantes, tiene un valor extremadamente bajo (5.44e-29 veces), lo que sugiere que esta variable es prácticamente irrelevante en el modelo. Por consiguiente, PROB_DESERCION cuando el resto de las variables se mantienen constantes, es extremadamente alta, incrementando las probabilidades de ubicar un unto de venta deserto en 3.45e+25 veces, lo que indica que esta variable tiene un efecto muy importante.

Para finalizar, estar en la región costa se asocia con una reducción de las probabilidades de deserción, haciéndolas 0.93 veces menos probables, aunque este efecto no es estadísticamente significativo.

Ahora, continuando con el análisis del enfoque de Random Forest, que consiste en un conjunto de árboles de decisión diseñados para identificar las características más relevantes que permiten predecir si un Punto de Venta es deserto o no.

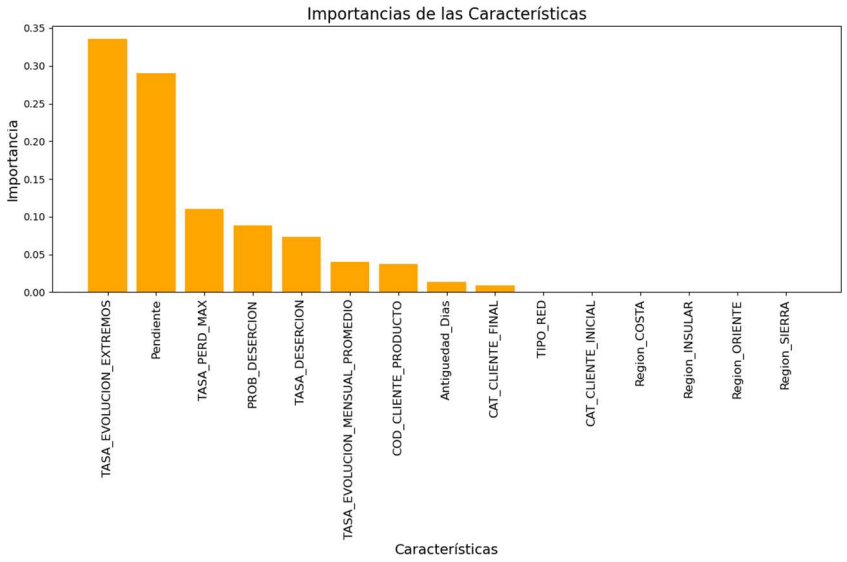

Figura 9 – Importancia de las Variables – Random Forest

El grafico nos muestra que la TASA_EVOLUCION_EXTREMOS es la más relevante para predecir la deserción, seguida de la Pendiente y la TASA_PERD_MAX. Esto quiere decir que el comportamiento de las ventas de los PDV son un aspecto importante en la deserción. También se destacan la PROB_DESERCION y la TASA_DESERCION, aunque su impacto es menor. Por otro lado, características como TIPO_RED y las Dummies de las regiones tienen una importancia de 0, lo que sugiere que no son útiles para el modelo.

Dentro del modelo de Random Forest, se aplicaron los hiperparámetros óptimos para evitar el sobreajuste.

Hiperparámetro	Valor	Descripción
n_estimators	39	Número de árboles en el bosque.
min_samples_split	10	Número mínimo de muestras requeridas para dividir un nodo.
max_leaf_nodes	18	Número máximo de nodos hoja. Limita la complejidad del modelo, lo que puede mejorar la interpretabilidad y reducir el sobreajuste.
max_features	0.5	Porcentaje de características consideradas para cada división.
max_depth	17	Profundidad máxima de los árboles. Permite capturar patrones complejos en los datos.
criterion	entropy	Método utilizado para medir la calidad de las divisiones.
bootstrap	True	Indica que se utiliza el muestreo bootstrap, aumentando la variabilidad entre los árboles y mejorando la capacidad de generalización del modelo.

Tabla 19 – Descripción Hiperparámetros Random Forest

Interpretación de resultados

Dentro del de estudio se realizó una comparación entre dos tipos de modelos: regresión logística y Random Forest. Al analizar los resultados de las pruebas, se concluye que el modelo más adecuado es el segundo, ya que presenta un mejor ajuste.

Métrica	Regresión Logística		Random Forest	
	Prueba	Entrenamiento	Prueba	Entrenamiento
Accuracy	0.9196	0.9194	0.9352	0.9362
Recall	0.9367	0.9372	0.9668	0.9691
Precision	0.9415	0.9425	0.9373	0.938
F1 Score	0.9391	0.9399	0.9518	0.9533
ROC AUC	0.9601	0.9574	0.9762	0.976

Tabla 20 – Resultados Pruebas - Regresión Logística vs Random Forest

En términos de exactitud (Accuracy), Random Forest es más preciso, ya que predice correctamente un mayor porcentaje de casos tanto en el conjunto de prueba como en el de entrenamiento.

Cuando se trata de la sensibilidad (Recall), que mide qué tan bien el modelo identifica los casos positivos, Random Forest también es superior, lo que significa que es mejor encontrando los casos que realmente deberían ser clasificados como desertores.

Sin embargo, en cuanto a la precisión, la Regresión Logística es ligeramente mejor. Esto quiere decir que, cuando predice un caso de deserción, tiene menos predicciones incorrectas. En cuanto al F1 Score, que combina precisión y sensibilidad en una sola métrica, Random Forest vuelve a ser mejor que la Regresión Logística. Esto muestra que Random Forest es más equilibrado entre encontrar los casos de deserción y evitar predicciones incorrectas.

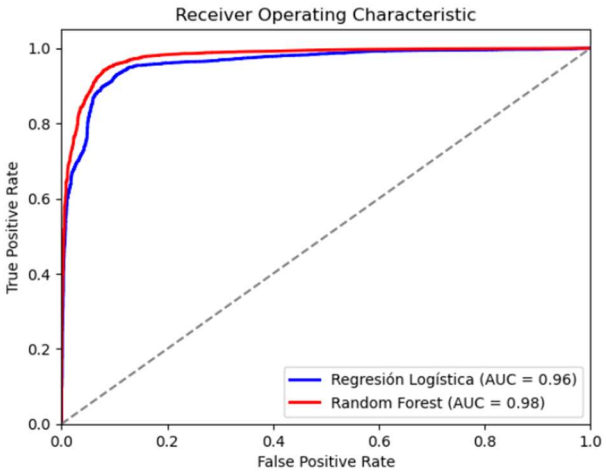

Figura 10 – Gráfico ROC – Regresión Logística vs Random Forest

En la curva ROC AUC, que mide qué tan bien el modelo distingue entre Desertor y No Desertor, Random Forest supera nuevamente a la Regresión Logística, lo que indica que discrimina mejor entre ambos tipos de casos.

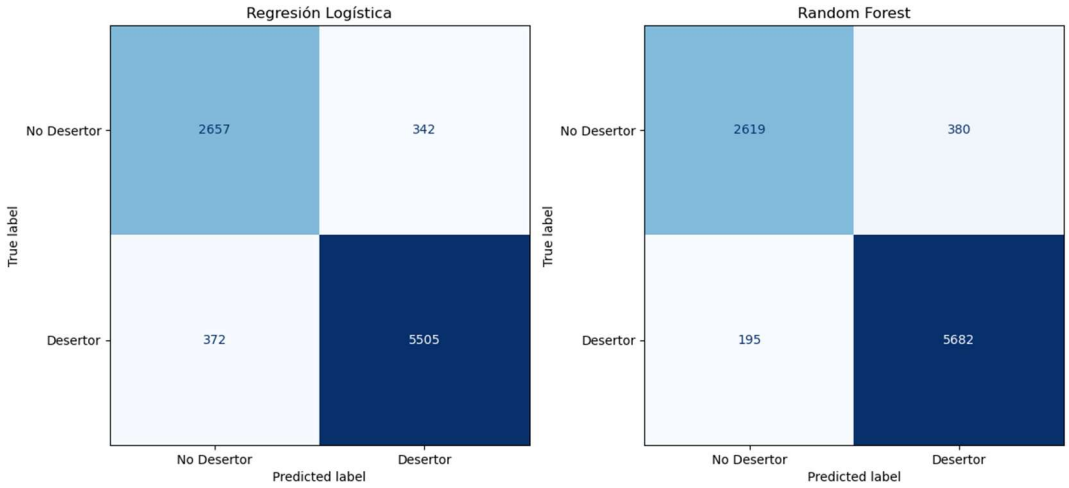

Figura 11 – Matriz de Confusión – Regresión Logística vs Random Forest

Si analizamos las matrices de confusión muestra que el modelo Regresión Logística identificó correctamente a 2657 personas que son no desertores, pero se equivocó con 342, clasificándolos como desertores por error. También clasificó correctamente a 5505 desertores, pero cometió 372 errores al considerar a algunos desertores como no desertores.

Consecuentemente, el modelo Random Forest identificó correctamente a 2619 no desertores y se equivocó con 380. Sin embargo, su mayor ventaja fue en los desertores, donde acertó con 5682 y solo se equivocó en 195 casos. Demostrando así que el modelo de Radom Forest es más eficaz para predecir a los Desertores.

Por otro lado, se procedió a analizar las diferentes métricas de ambos modelos, pero desde la perspectiva de las dos clases presentes:

Regresión Logística

Partición	Clase	Precisión	Recall	F1-Score	Accuracy
Prueba	Desertor	0.8772	0.8860	0.8816	0.9196
	No Desertor	0.9415	0.9367	0.9391	
Entrenamiento	Desertor	0.8730	0.8832	0.8781	0.9194
	No Desertor	0.9425	0.9372	0.9399	

Tabla 21 – Resultados Pruebas por Clase – Regresión Logística

Random Forest					
Partición	**Clase**	**Precisión**	**Recall**	**F1-Score**	**Accuracy**
Prueba	Desertor	0.9307	0.8733	0.9011	0.9352
	No Desertor	0.9373	0.9668	0.9518	
Entrenamiento	Desertor	0.9322	0.8690	0.8995	0.9362
	No Desertor	0.9380	0.9691	0.9533	

Tabla 22 – Resultados Pruebas por Clase – Random Forest

Desde esta perspectiva, podemos ratificar la fortaleza del modelo de Random Forest frente a la Regresión Logística, pues el rendimiento en cada una de las pruebas para cada una de las clases (Desertor y No Desertor es superior).

A pesar de que ambos modelos tienen un rendimiento bastante alto, Random Forest presenta mejores resultados en las pruebas, en especial al momento de predecir la clase “Desertor”, que, para efectos del presente trabajo, es nuestra variable objetivo. Para poner en contexto la información obtenida podemos tomar como referencia la métrica de precisión, en la cual Random Forest tiene un rendimiento de seis puntos porcentuales por encima de regresión Logística; el mismo comportamiento – aunque con porcentajes diferentes - lo tenemos presente en las pruebas de F1-Score y Accuracy.

Análisis de Supervivencia Clientes

Como se mencionó anteriormente, no solo es crucial conocer si un cliente tiene probabilidad de desertar, sino también identificar en qué momento es probable que ocurra la deserción dentro de un rango de tiempo específico. En este contexto, el análisis de supervivencia de clientes es una herramienta sumamente importante. A través de este enfoque, podemos proyectar el comportamiento esperado de la cartera de clientes en términos de retención a lo largo de un periodo definido, que en el presente caso de estudio abarca 12 meses. Además, el análisis se realizó desde diferentes perspectivas, como la ubicación geográfica y la categoría de cliente, lo que permitió identificar la estructura de la cartera de clientes y el grado de fidelización o retención que la empresa posee.

Análisis de modelo estadístico

Para llevar a cabo el análisis de supervivencia, se utilizó el modelo Kaplan-Meier, que permite estimar la probabilidad de que un cliente continúe activo - realizando transacciones - con la empresa durante un periodo determinado.

Este modelo estadístico se basa en dos variables calculadas dentro del dataset. La primera es **time**, que representa la cantidad de veces que un cliente realizó transacciones durante los doce meses analizados. El valor de time varía entre cero (si el cliente no estuvo activo en ningún momento) y doce (si el cliente realizó transacciones en todos los meses del periodo de análisis).

La segunda variable es **event**, la cual clasifica a los clientes en dos categorías: i) 0 – si el cliente no presentó deserción durante los doce meses, y ii) 1 – si el cliente desertó al menos una vez en ese periodo. Es importante destacar dos aspectos: en primer lugar, el cálculo de event depende del valor de time, y en segundo lugar, el criterio de deserción fue definido de manera rigurosa para obtener una tasa de fidelización lo más precisa posible. No obstante, es importante mencionar que se pueden aplicar distintos enfoques al definir los criterios para las variables time y event, dependiendo de los objetivos del análisis.

El modelo se basó en los datos reales registrados durante los doce meses anteriores de cada cliente y, con esta información, proyectó la tasa de retención para los próximos doce meses.

El análisis de supervivencia de clientes se realizó desde tres perspectivas clave:

- En primer lugar, el modelo se aplicó a la totalidad de los datos para calcular la tasa de supervivencia general. Estos valores promedio proporcionan una visión global del comportamiento de la cartera de clientes.
- A continuación, se segmentó el análisis por ubicación geográfica, dividiendo los clientes en cuatro grupos: sierra, costa, oriente e insular. Esto permitió identificar la tasa de supervivencia específica de cada región.
- Finalmente, el mismo enfoque se aplicó según la categoría de clientes, clasificados en función de su nivel de ventas actual: C, B, A, AA, AAA y VIP.

Este análisis nos permite obtener una visión más detallada de la retención de clientes desde diferentes ángulos, lo que facilita la identificación de patrones específicos y estrategias de retención.

Interpretación de resultados

Como se lo comentó anteriormente, el primer análisis realizado fue sobre la totalidad de la base de clientes – sin realizar ningún tipo de segmentación -. Sobre esta data, podemos identificar que la tasa de supervivencia de clientes, luego de un año es del 60.76%, con una deserción promedio mensual del 3.27% (39.29% en un año).

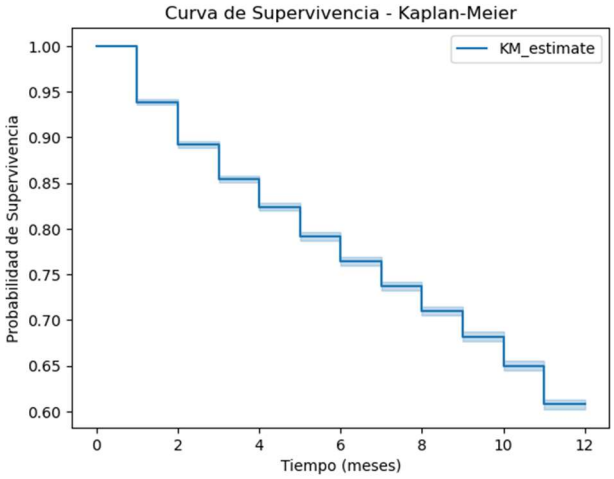

Figura 12 – Curva de Supervivencia de Clientes Global

Un resumen del pronóstico de supervivencia de clientes general se lo detalla a continuación:

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			32,002
1	93.88%	6.12%	1,959	30,043
2	89.24%	4.64%	1,484	28,559
3	85.43%	3.81%	1,220	27,339
4	82.39%	3.03%	971	26,368
5	79.18%	3.22%	1,029	25,339
6	76.44%	2.74%	876	24,463
7	73.76%	2.68%	859	23,604
8	71.00%	2.76%	882	22,722
9	68.18%	2.82%	904	21,818
10	64.97%	3.21%	1,027	20,791
11	60.76%	4.21%	1,346	19,445
12	60.76%	0.00%	-	19,445

Tabla 23 – Análisis Supervivencia Clientes Global

En un análisis más profundo, se procedió a segmentar la base de clientes en función de su ubicación geográfica por regiones en el territorio ecuatoriano: Costa, Sierra, Oriente e Insular. En este análisis podemos resaltar que la región que mejor comportamiento presenta luego de un año es la Insular, con una tasa de supervivencia del 71.15%, seguido de la Sierra (63.13%), y las regiones Costa y Oriente con un porcentaje muy parejo (57.27% y 55.66% respectivamente).

En la región Insular este comportamiento es esperado, dado el limitado espectro comercial que posee, los pocos clientes que la empresa capta en dicha región – 0.14% del total de la cartera – tienen una alta probabilidad de mantenerse activos, puesto que la presencia de competidores directos es muy limitada.

La región Sierra es otra región que presenta un rendimiento bastante alto, luego de un año la tasa de retención es del 63.13% - una deserción del 36.87% - principalmente se debe a que esta región, y sobre todo Pichincha, es la base de operaciones principales de la empresa (60% de los clientes se ubican aquí).

Finalmente tenemos a las regiones Costa y Oriente, con deserciones anuales del 42.73% y 44.34% respectivamente, si bien no son valores tan alarmantes, la supervivencia de clientes se explica por la poca presencia comercial que tenemos en dichos sectores.

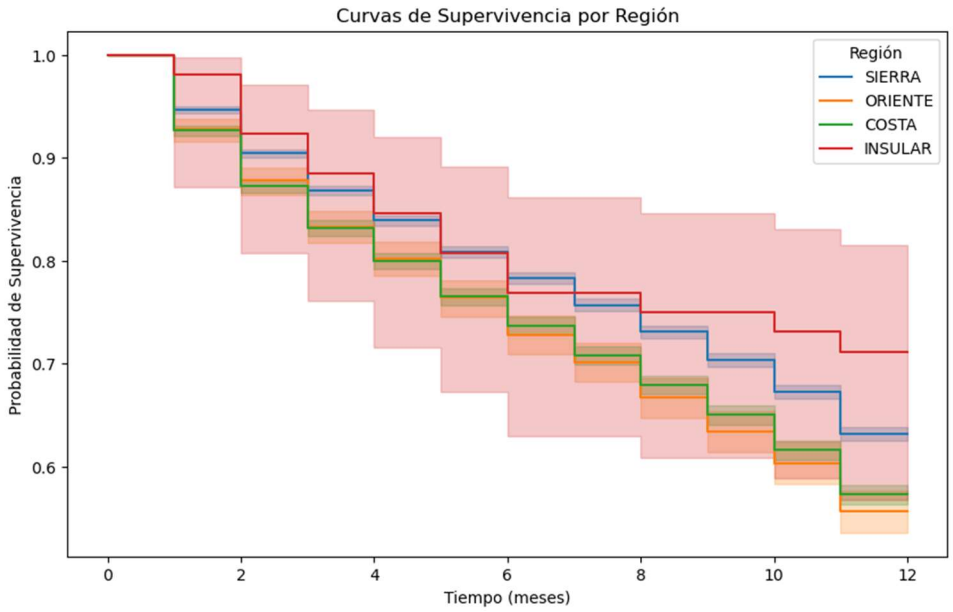

Figura 13 – Curva de Supervivencia de Clientes por Región

Un resumen del pronóstico de supervivencia de clientes en función de la región en la cual se ubican, se lo detalla a continuación:

MES	SIERRA	COSTA	ORIENTE	INSULAR
0	100.00%	100.00%	100.00%	100.00%
1	94.64%	92.64%	92.77%	98.08%
2	90.44%	87.24%	87.78%	92.31%
3	86.82%	83.19%	83.32%	88.46%
4	83.91%	79.94%	80.24%	84.62%
5	80.86%	76.54%	76.41%	80.77%
6	78.30%	73.66%	72.84%	76.92%
7	75.70%	70.79%	70.16%	76.92%
8	73.08%	67.93%	66.73%	75.00%
9	70.34%	65.04%	63.38%	75.00%
10	67.23%	61.58%	60.35%	73.08%
11	63.13%	57.27%	55.66%	71.15%
12	63.13%	57.27%	55.66%	71.15%

Tabla 24 – Análisis Supervivencia Clientes por Región

Un detalle pormenorizado de los resultados del análisis de la tasa de supervivencia de clientes por su región, se encuentra en la sección Anexos del presente trabajo.

Finalmente se procedió a realizar el análisis desde la perspectiva de categoría de clientes. En función del nivel de ventas que actualmente cada punto de venta posee, se lo ubica en una de las siguientes categorías (de menor a mayor) C, B, A, AA, AAA y VIP. Un punto importante a resaltar es la gran diferencia que existe entre las diferentes categorías, puesto que el segmento de clientes C, tiene una supervivencia del 53.78% luego de un año, mientras que el resto de segmentos promedian el 90%.

Además, si lo comparamos con la distribución del número de clientes por su segmento, la categoría C representa el 84% del total de clientes y el 21% del total de ventas. Sobre esta situación podemos concluir lo siguiente: i) este segmento de clientes es el que hace que las tasas de supervivencia calculadas anteriormente (general y por región) sean bajas y, ii) si bien es cierto que en apariencia la categoría C tiene un rendimiento preocupante, esto se explica debido a su bajo nivel transaccional, lo esporádico – cíclico – de sus ventas y la gran migración que existe hacia la competencia y su retorno. Sin perjuicio de aquello, el impacto que tiene este segmento en el resultado del negocio es moderado.

Por otro lado, tenemos a los segmentos B a VIP, que como se indicó anteriormente, presentan tasas de supervivencia elevadas (promedio del 90%),

esto se explica debido a que, en generalidad, un cliente va evolucionando sus ventas a lo largo del tiempo, esta evolución significa que el cliente ya posee cierto grado de fidelización hacia la empresa, por ejemplo, un cliente de los segmentos más altos (AAA y VIP) en promedio tienen una antigüedad de la relación comercial con la empresa más alta que un cliente tipo C, y esta antigüedad evidentemente trae consigo una mayor fidelización y consecuentemente mejores tasas de supervivencia.

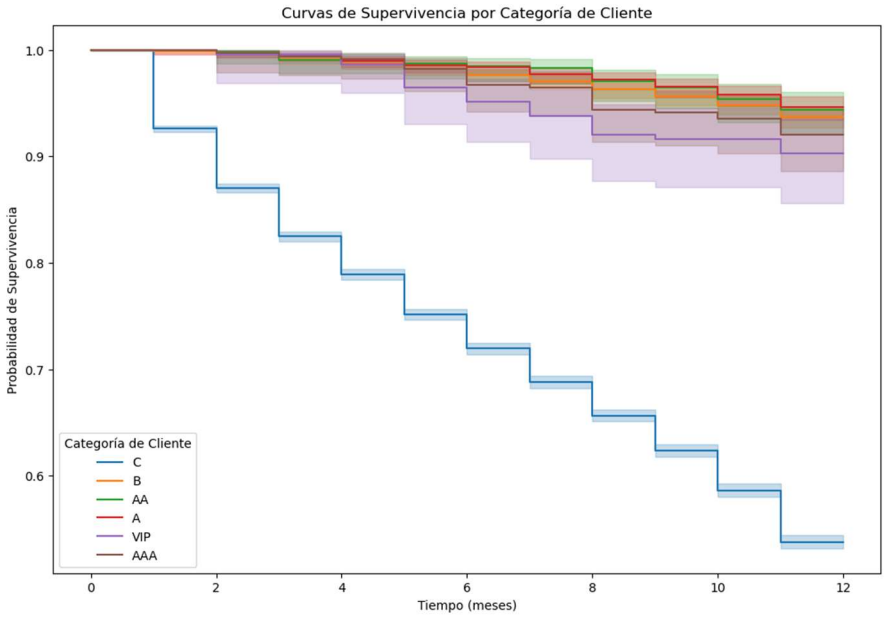

Figura 14 – Curva de Supervivencia por Categoría de Clientes

Un resumen del pronóstico de supervivencia de clientes en función de su segmento por nivel de ventas, se lo detalla a continuación:

MES	C	B	A	AA	AAA	VIP
0	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
1	92.60%	99.89%	99.94%	100.00%	100.00%	100.00%
2	87.04%	99.64%	99.70%	99.81%	99.71%	99.56%
3	82.50%	99.17%	99.47%	99.07%	99.41%	99.56%
4	78.92%	98.71%	99.00%	99.07%	99.12%	98.67%
5	75.14%	98.20%	98.58%	98.70%	98.23%	96.46%
6	71.93%	97.67%	98.35%	98.51%	96.76%	95.13%
7	68.81%	97.06%	97.70%	98.32%	96.46%	93.81%
8	65.65%	96.27%	97.23%	97.02%	94.40%	92.04%
9	62.36%	95.62%	96.52%	96.46%	94.10%	91.59%
10	58.63%	94.83%	95.81%	95.34%	93.51%	91.59%
11	53.78%	93.68%	94.63%	94.41%	92.04%	90.27%
12	53.78%	93.68%	94.63%	94.41%	92.04%	90.27%

Tabla 25 – Análisis Supervivencia por Categoría de Clientes

Un detalle pormenorizado de los resultados del análisis de la tasa de supervivencia de clientes por su segmento se encuentra en la sección Anexos del presente trabajo.

Implicaciones para la organización

Diseño de la estrategia

Una vez ejecutados los diferentes algoritmos que permiten identificar los potenciales clientes desertores, así como la probabilidad de que estos se mantengan activos en el futuro inmediato, se pudo obtener varias conclusiones relativas a la estructura de la cartera de clientes y su comportamiento en función de las diferentes segmentaciones que se realizaron.

Sin embargo, para que esta información sea verdaderamente útil, debe ser asimilada dentro de los diferentes procesos de la organización e incluirla como insumo para la toma de decisiones.

Como se mencionó al inicio, una de las actividades clave en la gestión comercial es la adecuada administración de los clientes.; no solo es necesario buscar nuevos clientes, sino también diseñar estrategias que permitan conservar y fidelizar los ya existentes. Es aquí donde los resultados del presente trabajo son de gran utilidad para este propósito.

En primer lugar, la información relacionada a los potenciales clientes desertores permite implementar un sistema de alerta temprana sobre los posibles puntos de venta con tendencia a la fuga, y sobre estos, articular estrategias comerciales tendientes a su retención. Además, el análisis de supervivencia ofrece una capa adicional al permitir una visión más amplia de la deserción, revelando cómo las probabilidades de abandono o permanencia varían significativamente entre segmentos. Específicamente, el sistema de alerta temprana debe priorizar a los tres segmentos de mayor categoría de clientes – VIP AAA AA – puesto que estos clientes son los de mayor valor para la empresa; sin perjuicio de aquello, el resto de los puntos de ventas no deberían ser desatendidos, sin embargo, las acciones a implementarse variarán en función del presupuesto y la capacidad de atención que la empresa tenga a su disposición.

La alerta temprana, busca advertir una tendencia hacia la deserción de estos puntos de venta; una vez identificados, se pueden ejecutar varias acciones encaminadas a su retención y fidelización. A continuación mencionamos algunas acciones que pueden ser implementadas.

Entre las acciones recomendadas, se puede iniciar con una campaña telefónica, por medio del departamento de atención al cliente, con el fin de, a manera de encuesta, identificar necesidades no atendidas o insatisfechas y proponer soluciones a las mismas. Este primer acercamiento, se lo puede masificar sobre

gran parte de la base de posibles desertores, puesto que el departamento de atención al cliente posee la capacidad instalada adecuada para asumir esta tarea (en términos físicos, humanos, tecnológicos, etc.), así también, el costo asociado a esta campaña es relativamente bajo, pues se utilizaría recursos que de todas formas la empresa tiene a su disposición.

En un segundo momento, es posible utilizar la fuerza de venta de campo para realizar visitas a los puntos de venta de mayor valor para la empresa – VIP AAA AA – identificados como posibles desertores, esta actividad, aunque simple, tiene un valor muy apreciado por el cliente, pues le brinda un sentido de cercanía con la empresa; en estas visitas no solo es posible conocer de primera mano la apreciación que tiene el cliente, sino también que permite dar soporte in situ, realizar la entrega de material publicitario, brandeo de locales, entre otras gestiones comerciales. La segmentación en este punto es importante, pues el personal de la fuerza de ventas no es lo suficientemente numeroso como para cubrir la totalidad de los clientes marcados como posibles desertores, de esta forma, se deberán seleccionar no solo los puntos de venta de mayor interés por su categoría de ventas, sino también por su ubicación geográfica, pues de esta manera se pueden establecer rutas o zonas de visitas en función de un calendario de trabajo.

Estas dos primeras estrategias podrían resolver casos de deserción debidos a factores operativos, como el olvido de credenciales o dificultades en el uso de la plataforma. Sin embargo, para los clientes que migran hacia la competencia por razones económicas, será necesario un enfoque distinto.

Para afrontar este problema, es necesario realizar un análisis pormenorizado del tipo de cliente con el cual estamos tratando, a fin de identificar el valor que este posee para la empresa, y en función de aquello ofrecer servicios diferenciados y reservados para un segmento específico de la cartera de clientes.

El negocio en el cual se desenvuelve la empresa objeto de este estudio se base en esquemas comisionales, sobre el cual la ganancia, tanto de la empresa como de cada uno de sus puntos de venta – clientes – gira en torno al volumen transaccional. En este sentido, la competencia ataca a los clientes ofreciendo ligeros cambios comisionales. Con el fin de responder a estas acciones, se puede ofrecer desde mejoras en las escalas comisionales, hasta líneas de crédito para cubrir la operación diaria de dichos comercios. Evidentemente, esto se lo debe realizar en función de la segmentación de la cartera, un cliente VIP o AAA serán perfiles de especial atención y sobre los cuales se puede contemplar estas opciones.

Por otro lado, para los clientes de categorías más bajas, como C o B, cuyo impacto económico es menor, se pueden resaltar los diferenciadores que ofrece la empresa, tales como un catálogo de productos más amplio, la estabilidad del servicio, atención al cliente, o programas de recompensas.

En resumen, la estrategia comercial debe estar alineada con diversos factores: la naturaleza del cliente, las necesidades no cubiertas, las implicaciones financieras de su posible fuga, la competencia y los objetivos estratégicos de la empresa, incluyendo la región y el nivel de ventas de cada cliente.

Implicaciones sobre Innovación

Dado que este proceso tiene un alto potencial y es relativamente novedoso para la empresa, es fundamental considerar los costos operativos asociados al desarrollo e implementación en un ambiente de producción, lo cual podría incluir la adquisición de software, almacenamiento de datos y, posiblemente, la contratación de expertos para garantizar una implementación efectiva.

Esta inversión permitirá el monitoreo continuo de los puntos de venta (PDV), lo que proporcionará una supervisión más eficiente y proactiva. La capacidad de detectar tempranamente las tendencias de deserción facilitará la toma de decisiones estratégicas para la retención de clientes, optimizando así la gestión comercial de la empresa.

La aplicación de este tipo de modelos no solo mejorará el uso de recursos y reducirá el margen de error humano, sino que también aumentará la capacidad de la empresa para retener clientes frente a los intentos de captura por parte de la competencia.

El verdadero valor agregado de esta implementación radica en la posibilidad de ofrecer una gestión proactiva y personalizada de la cartera de clientes. Al anticipar la deserción, la empresa podrá enfocar sus esfuerzos en estrategias de retención más efectivas, que no solo mitiguen la pérdida de clientes, sino que también fortalezcan las relaciones con los puntos de venta. Este enfoque permite a la empresa responder ágilmente a las demandas del mercado, brindando soluciones diferenciadas que satisfacen las expectativas de los clientes.

Sin embargo, para que esta propuesta pueda ofrecer los máximos beneficios posibles, es necesario integrarla de forma eficiente al resto de procesos que actualmente posee la empresa. Esta inclusión garantizará que los resultados que se obtengan de los algoritmos, puedan ser rápidamente incluida en la toma de decisiones y articulación de estrategias.

Además, la capacidad de replicar el modelo con nuevos datos asegura una adaptación continua a las dinámicas del mercado, generando un valor sostenible a largo plazo. Con esta implementación de modelos de machine learning y un enfoque centrado en la retención, la empresa no solo optimiza sus recursos, sino que también se posiciona estratégicamente en un mercado competitivo, impulsando su crecimiento y rentabilidad de manera consistente.

CONCLUSIONES

El presente trabajo tuvo como objetivo abordar la problemática de deserción de clientes dentro de una empresa auxiliar de servicios financieros. Dadas las condiciones particulares del mercado en el que se desenvuelve, la fuga de clientes es una constante a ser considerada dentro de la estrategia comercial empresarial.

Bajo las condiciones actuales en las cuales opera la empresa, y la envergadura de sus operaciones, el volumen de información relacionado con sus clientes y las transacciones que estos realizan son demasiado grandes como para trabajar con ellos con métodos tradicionales. Es aquí donde el Big Data y los algoritmos de machine Learning son de gran utilidad para proporcionar una respuesta altamente eficiente para solucionar los desafíos de análisis de información.

Los algoritmos que fueron usados para realizar el adecuado análisis de los posibles clientes desertores fueron regresión logística y random forest, mismos que fueron comparados entre sí con el fin de evaluar su desempeño e identificar el modelo que mejor se ajuste a la realidad del negocio. Así también, se aplicó un algoritmo de análisis de supervivencia - modelo Kaplan-Meier – para profundizar el análisis, dando una perspectiva más amplia respecto del comportamiento de la deserción, no solo desde una perspectiva temporal de doce meses, sino también por medio de un análisis por segmentación de clientes (categoría de ventas y ubicación geográfica).

Con el resultado de estos análisis, es posible articular estrategias comerciales enfocadas en la retención de los clientes potenciales de fuga, mejorando de esta manera la calidad de la cartera de clientes, sus indicadores de supervivencia, la fidelización de estos hacia los servicios que ofrece la empresa y consecuentemente, el desempeño económico y financiero de la organización.

Finalmente, es importante resaltar que los modelos elaborados en el presente trabajo deben integrarse adecuadamente a los procesos de la organización para que rindan los resultados esperados; desde la captura oportuna de la información, que sirve de insumo para correr los algoritmos – hasta el uso de los resultados obtenidos en las estrategias comerciales; es necesario lograr una sinergia entre departamentos con el propósito de aprovechar al máximo los beneficios que estos modelos pueden proporcionar.

RECOMENDACIONES

Para optimizar la retención de clientes en la empresa se recomienda la integración de los modelos desarrollados en un ambiente de producción dentro de los sistemas existentes. Lo cual ayuda a analizar los datos de forma oportuna y permitirá que los algoritmos operen de manera eficiente, facilitando el flujo de información entre los diferentes departamentos.

Es importante proporcionar capacitación continua al personal encargado en el uso de Big Data y algoritmos de machine learning. Esta formación mejorará la comprensión de las herramientas disponibles y fomentará un enfoque analítico en la toma de decisiones, lo cual ayudará a mejorar y a entender los beneficios de la adopción de estas nuevas tecnologías.

Puesto que las necesidades pueden cambiar debido a las condiciones del mercado, se sugiere establecer un proceso de monitoreo y actualización regular de los modelos de análisis. Esto con el fin garantizar que las predicciones y estrategias se mantengan acorde con las nuevas problemáticas que podrían surgir y así adaptarse a nuevas tendencias y desafíos.

Fomentar una cultura de análisis continuo de la información es otro aspecto fundamental. Establecer procesos para evaluar regularmente los datos de clientes y transacciones ayudará a detectar tendencias y oportunidades para mejorar la retención, asegurando que la empresa se mantenga competitiva.

Se recomienda el trabajo en equipo entre los diferentes departamentos de la empresa, como el área comercial, atención al cliente y análisis de información. Ya que será un punto importante para implementar eficazmente las estrategias de retención y optimizar el uso de los resultados del análisis. Al establecer métricas y KPIs claros, la empresa podrá evaluar y dar seguimiento a la efectividad de las estrategias implementadas, ajustando los pasos a seguir según sea necesario para maximizar el impacto en la retención de clientes.

BIBLIOGRAFÍA

Alaminos, A. (2023). *Árboles de decisión en R con Random Forest.* Alicante: Limencop. Obtenido de https://rua.ua.es/dspace/bitstream/10045/133067/1/Random_Forest_en_la_Investigacion_Social.pdf

Arana, C. (2021). *Modelos de aprendizaje automático mediante árboles de decisión.* Buenos Aires, Argentina.

Bahamonde Morales, D. I., & Tapia Pisarro, W. S. (2022). *Análisis comparativo del rendimiento de algoritmos de clasificación binaria en un conjunto de datos desbalanceados.* Quito.

Beltrán, C., & Barbona, I. (2022). Una evaluación del desempeño en la clasificación binaria mediante simulación: Árboles de clasificación y Bosques aleatorios. *REVISTA DE EPISTEMOLOGÍA Y CIENCIAS HUMANAS*. Obtenido de http://hdl.handle.net/2133/24321

Borja-Robalino, R., Monleón-Getino, A., & Rodellar, J. (2020). Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 184-196.

Gutiérrez González, D. (2020). *Técnicas de machine learning en el análisis del churn rate.*

Hernández, S. D. (2020). La Fidelización del Cliente y Retención del Cliente: Tendencia que se Exige Hoy en Día. *Gestión en el tercer milenio*, p. 5-13.

Javier, E.-Z. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, 1-16. doi:https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Lee, I. (2019). Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, 293-303.

Maisueche Cuadrado, A. (2019). *UTILIZACIÓN DEL MACHINE LEARNING EN LA INDUSTRIA 4.0.* UNIVERSIDAD DE VALLADOLID.

Martínez Fernández, T. C. (2022). *Comparación de modelos Machine Learning aplicados al riesgo de crédito.* Chile.

Navas Ayala, J. R. (2023). *Diseño de un modelo predictivo de fuga de clientes utilizando algoritmos de machine learning.*

Orenes Casanova, Y. (2022). *Contribuciones al problema de clasificación en machine learning.*

Paz, R. (2022). Application of Survival Analysis Techniques to the Study of Student Delay in the Civil Engineering Major-FACET-UNT. Obtenido de https://www.researchgate.net/publication/362346131_Aplicacion_de_Tecnicas_de_Analisis_de_Supervivencia_al_Estudio_del_Retraso_Estudiantil_en_la_Carrera_de_Ingenieria_Civil-FACET-UNT/fulltext/63e8325b6425237563a6d1e8/Aplicacion-de-Tecnicas-de-Analisis-de-S

Pinto Galindo, D. A. (2020). *Diseño de un Modelo Predictivo de Fuga de Clientes Utilizando Algoritmos Machine Learning.*

Ramirez, D. H. (2019). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD.* Pereira: Universidad Libre Seccional Pereira.

Rivero, F. (2022). Decision Tree in Machine Learning. *Varianza*, 39-46. Obtenido de https://ojs.umsa.bo/ojs/index.php/revistavarianza/article/download/433/365/595

Sánchez, A. (2021). *MATHEMATICAL MODELS IN SURVIVAL ANALYSIS.* Salamanca. Obtenido de https://gredos.usal.es/bitstream/handle/10366/150196/Memoria%20tfg_Andrea_S%C3%A1nchez_Moreno_Estad%C3%ADstica.pdf?sequence=1&isAllowed=y

Serra, A. (2020). Comparación de algoritmos de clasificación supervisada. *Escola Tècnica Superior d'Enginyeria Industrial de Barcelona*. Obtenido de https://upcommons.upc.edu/bitstream/handle/2117/330482/tfm-mueo-alexandre-serra.pdf

Shalev-Shwartz, S., & Ben-David, S. (2019). *Understanding machine learning: from theory to algorithms.* Nueva York: Cambridge University Press.

Sierchuk, S. (2022). *Predicción de Churn en Fintech: Una estrategia de retención integradora que utiliza algoritmos de Machine Learning con el objetivo de eficientizar el uso del presupuesto de Marketing.*

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295-316. doi:https://doi.org/10.1016/j.neucom.2020.07.061

ANEXOS

ANEXO 1 - SUPERVIVENCIA CLIENTES - REGION SIERRA

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			19,563
1	94.64%	5.36%	1,049	18,514
2	90.44%	4.20%	822	17,692
3	86.82%	3.61%	707	16,985
4	83.91%	2.91%	570	16,415
5	80.86%	3.05%	596	15,819
6	78.30%	2.57%	502	15,317
7	75.70%	2.60%	508	14,809
8	73.08%	2.62%	513	14,296
9	70.34%	2.74%	536	13,760
10	67.23%	3.10%	607	13,153
11	63.13%	4.10%	803	12,350
12	63.13%	0.00%	-	12,350

ANEXO 2 - SUPERVIVENCIA CLIENTES - REGION ORIENTE

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			2,242
1	92.77%	7.23%	162	2,080
2	87.78%	5.00%	112	1,968
3	83.32%	4.46%	100	1,868
4	80.24%	3.08%	69	1,799
5	76.41%	3.84%	86	1,713
6	72.84%	3.57%	80	1,633
7	70.16%	2.68%	60	1,573
8	66.73%	3.43%	77	1,496
9	63.38%	3.35%	75	1,421
10	60.35%	3.03%	68	1,353
11	55.66%	4.68%	105	1,248
12	55.66%	0.00%	-	1,248

ANEXO 3 - SUPERVIVENCIA CLIENTES - REGION COSTA

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			10,145
1	92.64%	7.36%	747	9,398
2	87.24%	5.39%	547	8,851
3	83.19%	4.05%	411	8,440
4	79.94%	3.25%	330	8,110
5	76.54%	3.40%	345	7,765
6	73.66%	2.88%	292	7,473
7	70.79%	2.87%	291	7,182
8	67.93%	2.87%	291	6,891
9	65.04%	2.89%	293	6,598
10	61.58%	3.46%	351	6,247
11	57.27%	4.31%	437	5,810
12	57.27%	0.00%	-	5,810

ANEXO 4 - SUPERVIVENCIA CLIENTES - REGION INSULAR

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			52
1	98.08%	1.92%	1	51
2	92.31%	5.77%	3	48
3	88.46%	3.85%	2	46
4	84.62%	3.85%	2	44
5	80.77%	3.85%	2	42
6	76.92%	3.85%	2	40
7	76.92%	0.00%	-	40
8	75.00%	1.92%	1	39
9	75.00%	0.00%	-	39
10	73.08%	1.92%	1	38
11	71.15%	1.92%	1	37
12	71.15%	0.00%	-	37

ANEXO 5 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO C

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			26,421
1	92.60%	7.40%	1,955	24,466
2	87.04%	5.56%	1,470	22,996
3	82.50%	4.53%	1,198	21,798
4	78.92%	3.58%	947	20,851
5	75.14%	3.78%	998	19,853
6	71.93%	3.21%	848	19,005
7	68.81%	3.13%	826	18,179
8	65.65%	3.16%	834	17,345
9	62.36%	3.29%	869	16,476
10	58.63%	3.73%	985	15,491
11	53.78%	4.85%	1,281	14,210
12	53.78%	0.00%	-	14,210

ANEXO 6 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO B

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			2,785
1	99.89%	0.11%	3	2,782
2	99.64%	0.25%	7	2,775
3	99.17%	0.47%	13	2,762
4	98.71%	0.47%	13	2,749
5	98.20%	0.50%	14	2,735
6	97.67%	0.54%	15	2,720
7	97.06%	0.61%	17	2,703
8	96.27%	0.79%	22	2,681
9	95.62%	0.65%	18	2,663
10	94.83%	0.79%	22	2,641
11	93.68%	1.15%	32	2,609
12	93.68%	0.00%	-	2,609

ANEXO 7 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO A

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			1,694
1	99.94%	0.06%	1	1,693
2	99.70%	0.24%	4	1,689
3	99.47%	0.24%	4	1,685
4	99.00%	0.47%	8	1,677
5	98.58%	0.41%	7	1,670
6	98.35%	0.24%	4	1,666
7	97.70%	0.65%	11	1,655
8	97.23%	0.47%	8	1,647
9	96.52%	0.71%	12	1,635
10	95.81%	0.71%	12	1,623
11	94.63%	1.18%	20	1,603
12	94.63%	0.00%	-	1,603

ANEXO 8 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO AA

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			537
1	100.00%	0.00%	-	537
2	99.81%	0.19%	1	536
3	99.07%	0.74%	4	532
4	99.07%	0.00%	-	532
5	98.70%	0.37%	2	530
6	98.51%	0.19%	1	529
7	98.32%	0.19%	1	528
8	97.02%	1.30%	7	521
9	96.46%	0.56%	3	518
10	95.34%	1.12%	6	512
11	94.41%	0.93%	5	507
12	94.41%	0.00%	-	507

ANEXO 9 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO AAA

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			339
1	100.00%	0.00%	-	339
2	99.71%	0.30%	1	338
3	99.41%	0.30%	1	337
4	99.12%	0.30%	1	336
5	98.23%	0.88%	3	333
6	96.76%	1.47%	5	328
7	96.46%	0.30%	1	327
8	94.40%	2.06%	7	320
9	94.10%	0.30%	1	319
10	93.51%	0.59%	2	317
11	92.04%	1.47%	5	312
12	92.04%	0.00%	-	312

ANEXO 10 - SUPERVIVENCIA CLIENTES - CLIENTES TIPO VIP

MES	RETENCION	DESERCION	CANTIDAD DESERTORES	CANTIDAD CLIENTES
0	100.00%			226
1	100.00%	0.00%	-	226
2	99.56%	0.44%	1	225
3	99.56%	0.00%	-	225
4	98.67%	0.88%	2	223
5	96.46%	2.21%	5	218
6	95.13%	1.33%	3	215
7	93.81%	1.33%	3	212
8	92.04%	1.77%	4	208
9	91.59%	0.44%	1	207
10	91.59%	0.00%	-	207
11	90.27%	1.33%	3	204
12	90.27%	0.00%	-	204