ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Caracterización del abandono de clientes mediante el uso de algoritmos de aprendizaje automático para empresas de firma electrónica en Ecuador.

PROYECTO INTEGRADOR

Previo la obtención del Título de:

MAGISTER EN CIENCIA DE DATOS

Presentado por:

Anel Ivette Martínez Chávez

Marlon Alexander Segarra Zambrano

GUAYAQUIL - ECUADOR

Año: 2024

Declaración Expresa

Nosotros Marlon Alexander Segarra Zambrano y Anel Ivette Martínez Chávez acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 13 de diciembre del 2024.

Firmado electrónicamente por:
ANEL IVETTE MARTINEZ CHAVEZ

Anel Martínez

MARLON ALEXANDER SEGARRA ZAMBRANO
Firmado digitalmente por MARLON ALEXANDER SEGARRA ZAMBRANO
Fecha: 2024.12.13 15:40:00 -05'00'

Marlon Segarra

Evaluadores

Firmado electrónicamente por:
ROMEO ELIAS CABRERA AREVALO

Romeo Cabrera

PROFESOR TUTOR

Firmado electrónicamente por:
ALLAN ROBERTO AVENDANO SUDARIO

Allan Avendaño

PROFESOR EVALUADOR

RESUMEN

En la actualidad el uso de firma electrónica para los distinto trámites está ganando mucha notoriedad razón por la cual hay cada vez más empresas que se dedican a este fin, lo cual incrementa la competencia entre estas y obliga a adoptar estrategias de fidelización o retención de clientes.

Es importante tener en cuenta las puntuaciones tanto de atención y recomendación de estos, así como el departamento que atiende a las personas y la cantidad de soportes que tienen, mediante el análisis de estos factores se pueden implementar promociones atractivas para los clientes.

Mientras más enfocado sea el público objetivo de estas promociones, más clientes terminarán prefiriendo nuestro producto en lugar de cambiarse a la competencia.

A diferencia del proceso tradicional que incluye contactar a todos los clientes sin ofrecer descuentos ni promociones a clientes en peligro de abandonar, este trabajo incluye el entrenamiento de un modelo de aprendizaje automático que nos ayude a predecir el abandono de clientes, además de obtener un segmento especifico de aquellos clientes con mayor riesgo de abandono y un dashboard interactivo donde se puede consultar la operatividad en cuanto a la atención al cliente.

Como resultado se pudo obtener un sistema de detección temprana de clientes con mayor probabilidad de abandono, obteniendo el top 10% el cual nos proporciona un beneficio económico del 63% en relación con una muestra aleatoria de clientes

Palabras clave: Firma electrónica, clientes, detección temprana, aprendizaje automático, deserción, promociones.

ABSTRACT

Currently, the use of electronic signatures for various procedures is gaining significant prominence. As a result, there is an increasing number of companies dedicated to this purpose, intensifying competition and prompting the adoption of customer loyalty or retention strategies.

It is crucial to consider customer scores in areas such as service quality and recommendations, as well as the department handling their requests and the number of support interactions. By analyzing these factors, it is possible to implement attractive promotions tailored to customer needs.

The more precisely these promotions are aimed at the target audience, the more likely customers will choose our product instead of opting for the competition. Unlike traditional processes, which involve contacting all customers without offering discounts or promotions to those at risk of leaving, this work incorporates training a machine learning model to predict customer churn. It also identifies a specific segment of customers with the highest risk of churn and provides an interactive dashboard to monitor customer service operations.

As a result, we developed an early detection system for customers with a high likelihood of churn, focusing on the top 10% segment, which yields a 63% economic benefit compared to a random sample of customers.

Keywords: electronic signature, customers, early detection, machine learning, churn, promotions.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

1 Planteamiento de la problemática

1.1 Descripción del problema

La firma electrónica es un conjunto de datos asociados a un mensaje y que identifican a la persona que firma el documento; todo documento digital ya sea un Word, Excel, PDF o XML que contenga una firma electrónica tendrá el mismo efecto jurídico que una firma manuscrita y podrá ser usado como prueba en un proceso judicial, toda persona que ha firmado electrónicamente con un certificado emitido por una entidad de certificación acreditada por ARCOTEL (Agencia de Regulación y Control de las Telecomunicaciones) no puede negar su autoría.

Según la Ley de comercio electrónico, firmas electrónicas y mensajes de datos (Ley No. 2002-67) determina en el artículo 18 que: “Las firmas electrónicas tendrán duración indefinida. Podrán ser revocadas, anuladas o suspendidas de conformidad con lo que el reglamento a esta ley señale”, además de que en el artículo 23 que “Salvo acuerdo contractual, el plazo de validez de los certificados de firma electrónica será el establecido en el reglamento a esta Ley”.

La implementación de la firma electrónica en el país empezó con el decreto no. 813 (reglamento general a la ley orgánica para la transformación digital y audiovisual) en el artículo 38 y 39 nos indica que la firma electrónica será de uso obligatorio para los procesos y servicios que brindan las entidades debido a que estos documentos deben ser firmados electrónicamente en el software oficial del ente regulador de la transformación digital. De conformidad con el artículo 22 de la Ley Orgánica para la Transformación Digital y Audiovisual, las entidades del sector público y privado están

obligados a implementar y aceptar dentro de sus diferentes procesos, documentos que hayan sido firmados electrónicamente.

La entidad designada, con Oficio Nro. SNAP-SNADP-2016-000395-O, de fecha 20 de julio de 2016, estableció que “Los certificados de firma electrónica con clasificación de persona natural o física son de uso personal y sirven para todos los actos públicos y privados de su titular, por lo que, en cualquier caso, el valor del certificado de firma electrónica de persona natural o física deberá ser asumido por los servidores públicos adquirentes”.

La firma electrónica actualmente tiene varios usos como pueden ser: trámites judiciales, operaciones comerciales, contractuales, entre otros, por lo que el uso de la firma electrónica abarca varios campos en nuestro país, por lo existen entidades de certificación acreditadas por ARCOTEL las cuales se encargan de brindar firmas electrónicas con un costo variado dependiendo de la vigencia que necesita el cliente, además de validar los documentos que hayan sido firmados electrónicamente a través de cualquier software de validación, siempre y cuando este sea compatible con los certificados de firma electrónica emitidos.

Debido a la gran apertura que se ha generado en los acuerdos para la transformación digital y audiovisual en el país se ha originado un competitivo panorama empresarial, en donde no se puede subestimar el impacto de la pérdida de clientes. Las empresas se enfrentan a un incremento del mercado en donde deben competir ferozmente para mantener sus clientes actuales y ganar nuevos clientes frente a los proveedores emergentes en el mercado de firma electrónica en donde la reducción de costos y la intensa presión competitiva son los factores claves para que las empresas exploten plenamente sus bases de clientes existente. En consecuencia, se implementan campañas de retención de clientes para poder detectar a aquellos que disminuyen su lealtad hacia la empresa (Nicolas Glady, 2009).

Debido a que se presenta un incremento en la tasa de consumidores se generan nuevas empresas para la competencia las cuales se centran en brindar el servicio o producto a un menor costo captando una mayor concentración de nuevos clientes en el mercado, frente al servicio o producto que ofrece la empresa tradicional (Lalwani, 2022).

Cuando incrementa la competencia, se crea un efecto domino que repercute a todos los involucrados en el giro de negocio de firmas, lo cual afecta directamente al beneficio económico de la empresa involucrada en el mercado. Mientras las empresas tradicionales se centran en mantener los precios del mercado fijos y no generar una devaluación del valor del precio y calidad del producto brindado al cliente para no crear una desigualdad en los ingresos que produce la empresa en el mercado para recibir un beneficio.

Es evidente el decremento en los porcentajes de ganancia mensuales generados por la empresa tradicional en comparación con los valores de ganancia obtenidos en años anteriores frente a un menor número de competidores, esto difiere con las cifras actuales de ganancias generadas mostrando una pérdida de ingresos y un decremento de clientes para la adquisición o renovación de un nuevo producto (Jaishankar Ganesh, 2000).

Ante esta problemática se ha determinado la importancia para las empresas obtener un control de la pérdida de clientes, para determinar los factores principales que originan un decremento de la clientela que cambian de un proveedor a otro (Mozer, 2000).

1.2 Justificación

Actualmente debido a una variedad de factores las organizaciones comerciales deben tomar buenas decisiones basadas en lo que ahora se conoce como inteligencia de negocio. En las empresas la gestión de la deserción de cliente se ha vuelto un factor de

importancia ya que se ha demostrado que retener a los clientes antiguos puede generar más beneficios que conseguir nuevos consumidores ya que esto puede llevar mucho tiempo y ser más costoso tal como lo plantea (Glory Sam, 2024)

El presente trabajo busca identificar la posible deserción de los clientes en la empresa firma electrónica de forma temprana y poder generar una medida proactiva de retención con el objetivo de predecir una respuesta favorable de los clientes existentes para conservarlos. (Wouter Verbeke, 2011). Retener a los clientes existentes es tan importante como captar nuevos clientes, mediante la predicción de la pérdida de clientes permitirá a la empresa adoptar una mejor estrategia de producción. Usando esta probabilidad de abandono, se otorgará otorgando a los especialistas en marketing perfiles de clientes para la generación de campañas de reactivación más específicas para clientes en riesgo, dirigiéndolos a un segmento del mercado con mensajes y acciones enfocados a impulsar la adquisición y retención de estos clientes (Marcel Karnstedt, 2011).

Se busca analizar los datos recopilados de la empresa de los años 2022 al 2023 para generar una selección de muestra de datos relevantes que permitan identificar las variables de interés que se involucran directamente en la deserción de los clientes que poseen firma electrónica para el año siguiente.

El proceso actual que tiene la empresa para manejar la renovación de la firma es generar una campaña masiva mensual de los clientes que poseen firma electrónica por medio de WhatsApp o correo electrónico para comunicar del estado de la vigencia de la firma y solicitando la confirmación del cliente con la recepción de los datos de renovación y validación de identidad para poder renovar la firma. Pero en estas campañas generadas no se posee un registro oficial de los clientes que continúan con la empresa y de los que prefieren abandonar el servicio o ya lo ha abandonado.

La solución planteada es la creación de un modelo de aprendizaje automático el cual toma características claves de la atención al cliente como la consulta de solicitudes de compra del servicio y los soportes técnico postventa brindados para obtener una correcta probabilidad de que los clientes que decidan no renovar su firma electrónica y usando

esta probabilidad, segmentarlos en un grupo dedicado para producir una campaña preventiva de aquellos clientes que vayan a desistir del servicio, cuáles son los principales factores del posible abandono y de esta manera obtener un ahorro en el costo el envío de campañas masivas enfocadas a los clientes objetivos. Esta solución resuelve el problema de identificación de potencial abandono de clientes e identificación de factores determinantes del mismo (Praveen Lalwani, 2021).

1.3 Objetivos

1.3.1 Objetivo General

Caracterizar los clientes perdidos y retenidos mediante el cálculo de la probabilidad de abandono usando algoritmos de aprendizaje automático con la finalidad de aplicar estrategias de retención de clientes en la empresa.

1.3.2 Objetivos Específicos

1. Comparar el rendimiento de algoritmos de aprendizaje automático en la predicción de la probabilidad de deserción de clientes de firma electrónica, con la finalidad de escoger e implementar el mejor modelo.
2. Determinar las variables más influyentes en modelo de aprendizaje automático mediante el análisis de importancia de estas variables para determinar una correcta identificación de posibles acciones a tomar presentes en el proceso de atención al cliente.
3. Caracterizar a los clientes usando su probabilidad de abandono mediante el uso de un aplicativo para realizar una campaña preventiva de aquellos clientes que vayan a desistir del servicio.

1.4 Metodología

Para comenzar a implementar la solución propuesta existen cinco etapas para analizar y construir un modelo predictivo que será de vital importancia para la empresa es imperativo seguir los siguientes pasos:

1. Seleccionar muestras para su análisis.
2. Definir las variables explicativas para el modelo.
3. Limpieza y transformación de datos.
4. Construcción de los modelos.
5. Evaluación de los modelos.

Estos 5 pasos para el manejo del abandono de clientes fueron introducidos por primera vez en (Datta, 2000).

Los datos utilizados para el entrenamiento de nuestro modelo se basan en las encuestas de satisfacción realizados a los clientes en donde los datos obtenidos se evalúan en conjunto con la base de las ventas generadas en el portal que maneja la empresa. Además de relacionarlo con otra base de datos de soportes generados por los clientes que contrataron el servicio de firma electrónica, para poder obtener un nuevo conjunto de datos y producir una validación de los clientes que renuevan o no su servicio desde el año 2022 a 2023.Con la finalidad de realizar una correcta clasificación de clientes retenidos y perdidos en los años próximos.

La fase primordial en la ejecución de este proyecto se basa en el preprocesamiento de la data en la cual se consideran los siguientes aspectos:

1. La preparación de la data con las características que se considere de importancia para el análisis de nuestra problemática.
2. La limpieza de los datos para poder ejecutar de manera correcta los algoritmos de aprendizaje automático de forma que garantice la calidad de los datos transformando estos en un formato consistente y legible.
3. El etiquetado y la inclusión de nuevas métricas que nos va a permitir proporcionar contexto para que el modelo de machine learning pueda aprender de ellos.

Tal como se recomienda en (Seyed Hossein Iranmanesh, 2019)

El flujo completo del proceso de creación de nuestro modelo predictivo se evidencia en la figura 1.1 cuya estructura se basa en el estudio de (IRFAN ULLAH, 2019).

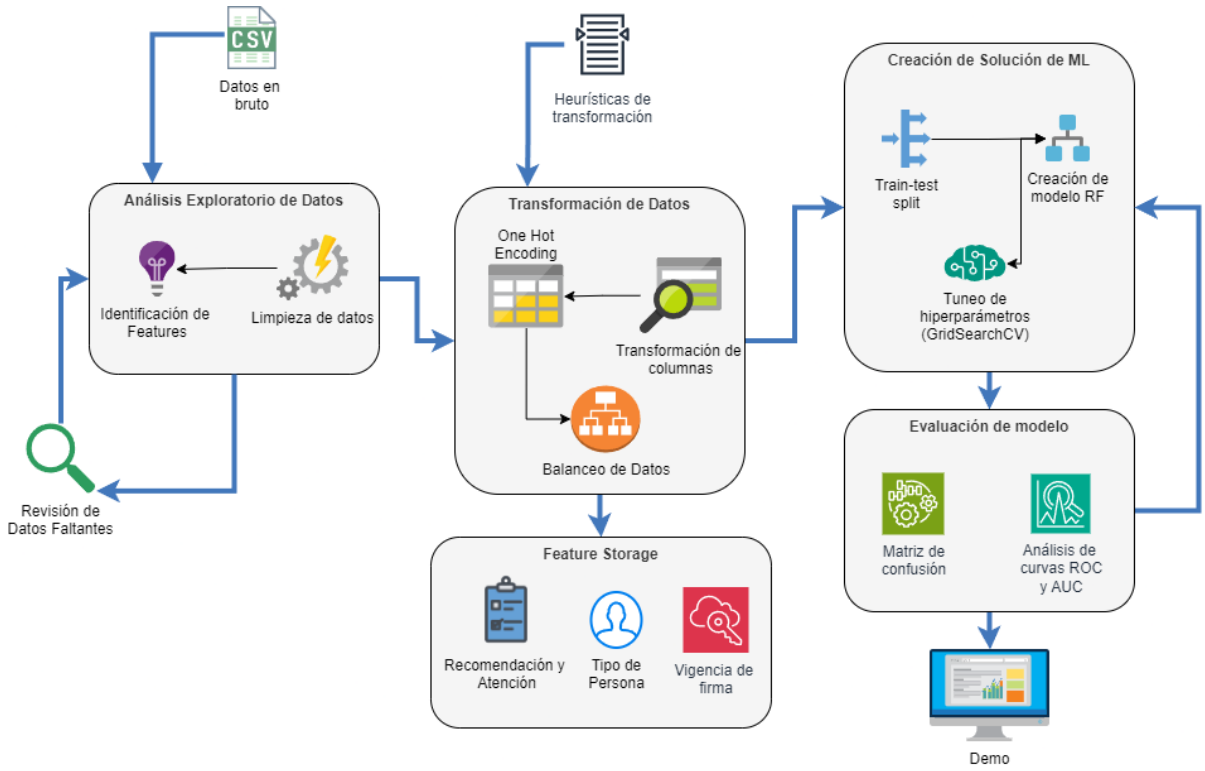

Figura 1.1:Flujo de creación del modelo

Los algoritmos con los que se realizará la comparación de su exactitud y precisión son: Decision Tree, Random Forest, Bagging, Catboost, XGboost, LightGBM y Stacking, es importante aclarar que estos algoritmos son los más usados en los estudios anteriores que se han realizado para la predicción del abandono de clientes tal como lo presenta (Anitha M A, 2022).

La importancia de las variables se realizará una vez se escoja el modelo que mejor exactitud y precisión posea mediante la función de “feature importances” del modelo. De

acuerdo con esto se identificará cuales características tienen el mayor peso al momento de realizar la predicción tal como se muestra en (Levent ÇALLI, 2022) y de esta forma revisar estas características en el aplicativo correspondiente con la finalidad de identificar los puntos de dolor en la atención al cliente, y con estos resultados tomar las decisiones recomendadas por los expertos en el área.

El aplicativo cumplirá la función de visualizar los datos suministrados mediante un reporte en Excel, el aplicativo contendrá 4 secciones las cuales serán:

- **Resumen ejecutivo:** se mostrará una proyección de los beneficios y pérfidas de las predicciones y una distribución de los clientes retenidos y desertores.
- **Análisis detallado:** un análisis de soportes mensuales con un filtro por equipo el cual gestiono ese requerimiento, además de un desglose de los clientes retenidos y desertores según la fuente de captación.
- **Proyecciones económicas:** se mostrará un total de predicciones por equipo y mes junto a sus beneficios y perdidas por predicciones.
- **Predicciones:** un resumen de ganancias y pérdidas que genera en la empresa, con una opción de descarga de reporte.

Este aplicativo sería utilizado por la gerencia, el departamento de operaciones y marketing para la reunión mensual con los accionistas con el objetivo de medir la calidad operativa de las renovaciones de firma.

Finalmente se utilizará la probabilidad de abandono del cliente junto con la importancia de las variables y mediante el uso de la estadística descriptiva se procederá a utilizar el conocimiento adquirido por los expertos en el negocio de firma electrónica para generar las respectivas estrategias preventivas ante el posible abandono de estos clientes.

1.5 Resultados Esperados

Con la aplicación de nuestro modelo de aprendizaje automático se espera que nuestra propuesta de solución identifique correctamente los clientes en riesgo de abandono para lograr una correcta orientación del público objetivo. Para finalmente generar un aplicativo

que permita identificar los factores clave que provocan el abandono de los clientes, de esta manera orientar una campaña de retención de clientes y así poder aumentar las ventas logrando un beneficio económico para la empresa, disminuyendo costos por campañas masivas generados para lograr una comunicación con el cliente cuya vigencia del servicio está por expirar.

Con el resultado obtenido del aplicativo, se busca brindar una herramienta de consulta para el departamento de márquetin y operaciones el cual se enfoca directamente en dos equipos:

- El equipo de renovaciones que se encarga de la venta del servicio.
- El de soporte que se encarga de brindar la asistencia técnica al cliente.

Para el departamento de márquetin este aplicativo le sirve para poder realizar un análisis de los posibles clientes que abandonarán el servicio y de esta forma proporcionar perspectivas para mejorar estrategias comerciales, personalizando campañas de marketing para fidelizar al cliente con la empresa.

Para el departamento de renovaciones le sirve para detectar las características que generan molestia en el cliente durante el proceso de renovación de una firma electrónica y de esta forma poder enfocarse a realizar una mejora en el proceso de venta con el cliente.

Para el departamento de soporte le sirve para mejorar la experiencia de los soportes técnicos brindados al cliente y de esta forma otorgar un servicio más rápido con un mayor compromiso con el cliente.

1.6 Dataset

Uno de los puntos más complejos de tratar con una dataset con una gran cantidad de datos es la comprensión de la semántica de los campos que pueden cambiar con el tiempo debido a eventos externos como campañas o promociones. Por ese motivo es importante conocer el contexto de estas variables de cambio para poder generar una correcta limpieza y tratamiento de los datos. (Richard J. Oentaryo, 2012).

Los datos utilizados para este estudio son desde el 2022 hasta el 2023 los cuales se dividen en 3 bases diferentes las cuales se detallan a continuación:

Base de datos de satisfacción de cliente: La base de datos adquirida es a partir de la encuesta de satisfacción y valoración del servicio obtenido que el cliente recibe al obtener la firma electrónica, en la cual se consulta al cliente el medio por el cual se enteró de nosotros, el nivel de satisfacción del servicio brindado durante el proceso de obtención y la consulta del nivel de recomendación del producto.

Base de datos de la transacción: La base de datos adquirida comprende el valor del producto, las fechas de registro y aprobación de la transacción, la ciudad, el medio de atención y el medio de contacto para la obtención de la firma electrónica, además de indicar si es cliente nuevo o desea renovación de su producto.

Base de datos de soporte: La base de datos adquirida comprende los soportes brindados al cliente mediante el canal digital WhatsApp y la página web en donde se puede identificar el tipo de soporte brindado, duración del soporte, departamento asignado al soporte, además de una encuesta de satisfacción brindada al cliente sobre el servicio.

El tamaño de todo el conjunto de datos antes del preprocesamiento es de: 698.618,00. Luego de aplicar el preprocesamiento y limpieza respectiva en donde se determinaron aquellos datos relevantes para nuestro estudio, debido a esto el nuevo tamaño del conjunto de datos es de 323.755 datos, en donde se explicará en el capítulo 3 las razones de la reducción del dataset.

A continuación, en la tabla 1.1, procedemos a detallar las características de los datos.

Características	Tipo de dato	Descripción
Tipo	Categórico	Tipo de persona (PN (persona natural), RL(representante legal), ME(miembro de empresa)).
Medio	Categórico	Medio de conocimiento de la empresa.
Atención	Categórico	Calificación de atención brindada.
Recomendación	Categórico	Recomendación a otros clientes.
Tipo de atención	Categórico	Se dividen en dos: validación en línea y citas express.
Vigencia	Categórico	Número de años de vigencia del servicio.
Valor del servicio	Numérico	Valor del servicio.
Provincia	Categórico	Provincia donde se atendió al cliente.
Renovación	Categórico	Identifica si el cliente viene por renovación.
Cliente retenido	Categórico	Etiqueta que indica si el cliente renovó en el siguiente año calendario.

Tabla 1.1:Descripción de las características de los datos

A continuación, en la tabla 1.2, presentamos una visualización tabular de los datos:

Tipo	Medio	Atención	Recomendación	Tipo de atención	Vigencia	Valor de servicio	provincia	renovación	Cliente retenido
PN	Google	5	10	Validación en línea	2	35.84	Pichincha	No	0
ME	Recomendación	4	5	Cita express	1	20.16	Guayas	No	0
RL	Era cliente	3	7	Validación en línea	3	30.25	Azuay	Si	1
PN	Redes sociales	4	9	Validación en línea	4	40.25	Machala	Si	1
ME	Google	2	4	Cita express	5	45.96	Guayas	Si	1

Tabla 1.2:Representación tabular de los datos.

Capítulo 2

2 Estado del arte

2.1 Soluciones al problema de abandono de clientes

2.1.1 Solución Heurística

La retención de clientes es crucial, ya que adquirir nuevos clientes es generalmente más costoso que mantener a los existentes. Una solución heurística efectiva para abordar este problema es mediante estrategias de contacto proactivo a través de redes sociales, enfocadas en clientes cuyo servicio está próximo a expirar usando una base de datos de clientes y actualizando constantemente de forma que se elimina del listado de clientes para campañas que no respondieron en la anterior.

Entre las posibles soluciones para esta problemática se pueden tomar en cuenta la siguiente solución:

- Integrar sistemas de CRM con herramientas de gestión de redes sociales para automatizar el proceso de identificación, contacto y seguimiento de clientes en riesgo.
- Utilizar chatbots y asistentes virtuales para proporcionar respuestas rápidas y personalizadas a consultas comunes, mejorando así la experiencia del cliente.

En las áreas de finanzas y marketing se orientan principalmente a la implementación de métodos estadísticos para evaluar la predicción de abandono de cliente, en el trabajo de determina el análisis de la prueba T y de Chi cuadrado en el cual analizan la inclinación al cambio en función de la satisfacción del cliente en dos grupos: la calidad técnica y calidad funcional (Mittal, 1998).

Además de otros estudios que utilizan el análisis de varianza (ANOVA) para examinar las diferencias en las intenciones de comportamiento de los clientes en función de la calidad del servicio percibida (Zeithaml, 1996).

2.1.2 Solución Basada en Redes Neuronales

En la predicción de abandono de cliente orientado a la industria de telecomunicaciones, donde se obtienen muchos datos, el uso de las redes neuronales para predecir este comportamiento es una de las mejores opciones, ya que brinda mejores resultados que los modelos tradicionales al tener 3 capas: la de entrada, la escondida y la de salida, además de asignar un peso aleatorio que mejora según se entrena el modelo y encontrar los pesos óptimos para obtener el mejor modelo tal como indica (Eria, 2018).

El artículo (S. Agrawal, 2018) aborda la predicción del abandono de clientes en el sector de telecomunicaciones mediante el análisis de patrones de comportamiento utilizando técnicas de deep learning. Emplea una red neuronal profunda multicapa para construir un modelo de clasificación no lineal. El modelo predice el abandono utilizando características del cliente, de soporte, de uso y contextuales. Este enfoque permite identificar de manera precisa los clientes con mayor riesgo de abandonar, ayudando a las empresas a implementar estrategias de retención más efectivas.

2.1.3 Solución Basada en Algoritmos de aprendizaje Supervisados

Para abordar esta problemática mediante el uso de algoritmos de aprendizaje supervisado podemos destacar el trabajo de (Louis Geiler, An effective strategy for churn prediction and customer profiling, 2022). En este estudio se utiliza una combinación de aprendizaje no supervisado para identificar únicamente patrones de comportamiento de estos clientes, sin embargo, la predicción se lleva a cabo usando algoritmos tales como: decision tree, random forest, regresión lineal y técnicas de ensamble tales como XGBoost además de realizar un balanceo de datos mediante la técnica del Undersampling dando como resultado final y modelo muy bueno usando XGBoost que obtuvo el mejor rendimiento.

Se puede destacar que en otros trabajos tales como (Louis Geiler, A survey on machine learning methods for churn prediction, 2022) presentan varios algoritmos de aprendizaje supervisado para predecir la deserción de clientes, estos implementan arboles de decisión, regresión logística, y métodos de ensamble: boosting, bagging y stacking. Con los cual se puede destacar el uso del Gradient Boosting Machine, el cual permite mejorar la precisión de la predicción, debido a que permite ajustar un nuevo árbol de decisión con los residuos del modelo anterior.

2.2 Fases de predicción del abandono de cliente

Para el estudio planteado se exponen diferentes factores que representan un reto en la tarea de predicción, los cuales representan en cada fase del estudio un factor primordial de revisión. A continuación, en la figura 2.1, indicaremos cuales son estos factores que se validan dentro de nuestro análisis de predicción con la data obtenida a partir de las bases entregadas por la empresa.

Figura 2.1:Limitantes del preprocesamiento de datos

Cantidad de datos obtenidos: De los datos obtenidos, se consideró para mejorar el estudio del modelo solo aplicar a la revisión de clientes que generan la renovación de la firma electrónica con vigencia de 1 a 2 años, para determinar con mayor precisión el análisis anual de renovación de la empresa para cada cliente.

Selección de los datos: se consideró a clientes principales que obtienen la firma electrónica por los medios de servicios de atención express o en línea, y no validando datos de clientes preferenciales, terceros vinculados o agentes que manejan lotes de compra de firmas electrónicas con valores atípicos (un solo pago de n firmas en una factura) y aplicaciones de descuentos para estos.

Semántica de la data: Para estudiar las variables presentadas en la data obtenida, es importante comprender el significado de cada variable que representa en los datos, en el caso de la vigencia y el campo de renovación que determina si el cliente es nuevo, se definió con valores numéricos en donde 0 es cliente nuevo y 1 es cliente que renueva el servicio de la firma. Este análisis nos brindó poder determinar la relación y el uso de estas variables en el estudio.

Entre estos 3 factores se desea equilibrar la cantidad y calidad de datos (selección y semántica de los datos) para construir un modelo robusto que nos permita extraer características relevantes y representativas del problema real del estudio como se indica en (Glory Sam, 2024). Y evitar en nuestro modelo un overfitting o underfitting evitando de esta forma cometer el tipico error de que el aumento del volumen de los datos involucra un mejor rendimiento del modelo, para finalmente evitar en un futuro un incremento de costos por recursos de almacenamiento y procesamiento.

Selección y validación del modelo: Para este estudio se evaluarán varios tipos de modelos los cuales se consideran los más relevantes para este tipo de análisis como son random forest, bagging y boosting. También se modificarán los respectivos hiperparámetros con la finalidad de obtener un rendimiento óptimo en la predicción de los clientes retenidos y perdidos.

De acuerdo con los factores antes mencionados los cuales representan un desafío en las fases del análisis de datos para nuestro modelo, mediante la literatura revisada, podemos determinar en la figura 2.2 las fases claves en el proceso de predicción del abandono de clientes para el negocio de firma electrónica las cuales son:

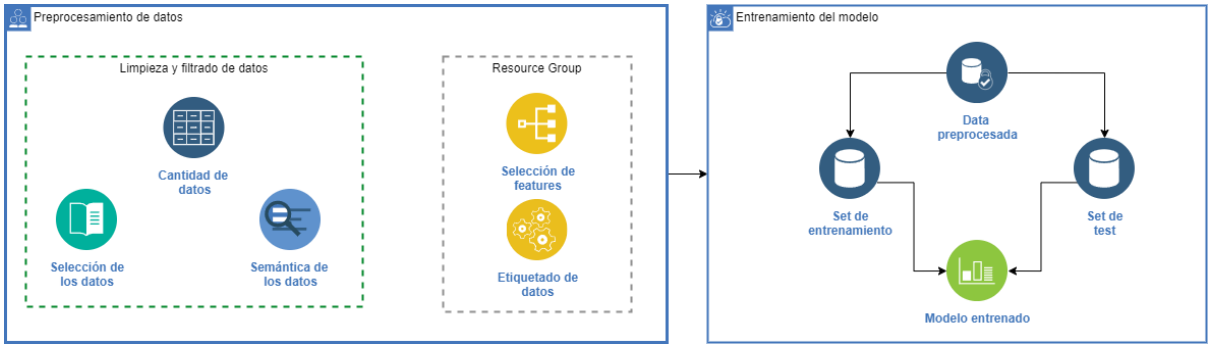

Figura 2.2:Fases del análisis de datos del modelo fuente: elaboración propia.

Selección de features: Para la selección de las columnas relevantes para nuestro estudio, se determinó una revisión en conjunto con el departamento de operaciones de la empresa para determinar cuáles son los factores relevantes que se consideran en las campañas de renovaciones generadas con el cliente, en donde se determinaron las siguientes: el valor de la firma, vigencia, tipo de atención (express o en línea), tipo de persona (persona natural, miembro de empresa o representante legal), sucursal de atención y si es o no un cliente nuevo.

Etiquetado de datos: Es una técnica que se basa en la transformación de la data categórica a una representación numérica, asignando de esta forma un número entero que representa la presencia o ausencia de la variable como lo indica (Kedar Potdar, 2017).

Para nuestro estudio se realizará el etiquetado de datos de los clientes retenidos siendo en este caso 1 la retención exitosa del cliente y 0 el abandono del cliente, además de la aplicación del One Hot Encoding a las variables categóricas presentes en los datos suministrados.

2.3 Definiciones y modelos

Las investigaciones relacionadas con el tema de predicción de abandono de clientes o Customer Churn Prediction aplican los modelos de Decision Tree, Logistic Regression, Random Forest,Boosting entre otros. A continuación, en la figura 2.3, presentamos los modelos más utilizados por los investigadores en esta problemática.

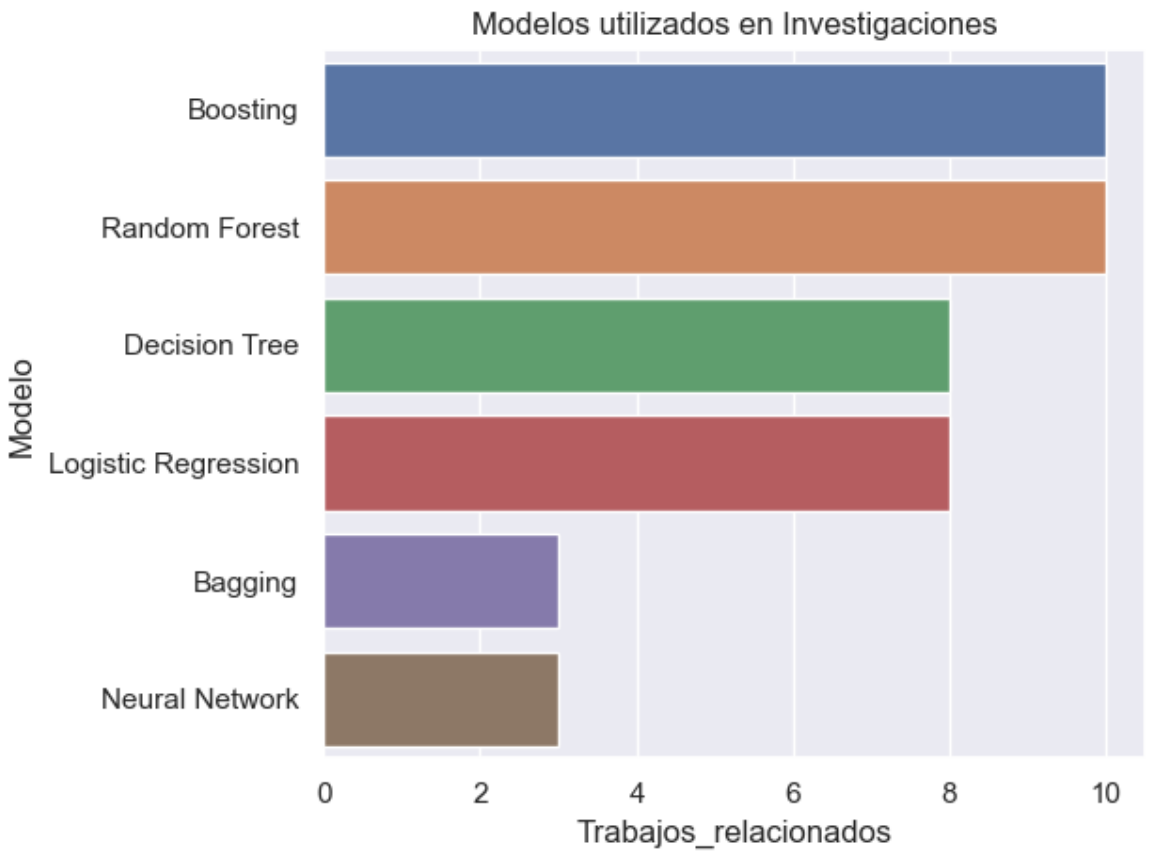

Figura 2.3:Modelos utilizados en investigaciones. Fuente: Elaboración propia.

Customer Churn

El abandono de clientes hace referencia a la terminación de la relación entre los clientes y la empresa proveedora del servicio lo cual reduce los ingresos y aumenta el costo de adquirir nuevos clientes. Por esta razón los métodos de predicción buscan mejorar la

eficiencia en cuanto a campañas de retención de clientes y maximizar las ganancias. (Ping Jiang, 2024)

Balanceo de datos

En el desarrollo de los distintos modelos de aprendizaje automático, el desbalanceo de la data es otro de los factores que generan dificultades en la distribución equilibrada de las clases en nuestros datos, y como la mayoría de los algoritmos de clasificación asumen que los datos de entrenamiento están equilibrados en distribución de clases, genera un decrecimiento del rendimiento del modelo en su clasificación. Por tal razón, en la figura 2.4, se indican las distintas técnicas para tratar con la data desbalanceada según (Mohammed, 2020).

Figura 2.4:DownSampling y Upsampling. Fuente: (Barros, 2019)

Oversampling: Se trata de un método de balanceo de datos el cual se basa en realizar réplicas aleatorias de los datos de la clase minoritaria e incorporarlas al dataset de forma que se finalice con la misma cantidad de datos que la clase mayoritaria.

El oversampling contribuye a evitar reducir la cantidad de datos en nuestro dataset, esto es de vital importancia para que nuestro modelo posea los datos suficientes para el entrenamiento, además de que nos permite mantener la información inicial que estará contenida dentro del conjunto de datos de remuestreo, sin embargo, la duplicidad de esta información puede generar un overfitting al momento de realizar nuestro modelo.

Undersampling: Se trata de un método de balanceo de datos el cual se basa en eliminar de forma aleatoria los datos de la clase mayoritaria de forma que se finalice con la misma cantidad de datos que la clase minoritaria.

El undersampling contribuye al balanceo de la data sin la necesidad de duplicar datos existentes manteniendo de esta forma los datos reales sin realizar cambio alguno o crear datos ficticios, de esta forma se asegura que nuestro modelo trabaje con datos reales con la desventaja de reducir el tamaño final del dataset y se perderá una cierta cantidad de información que serviría para el entrenamiento del modelo.

Decision Tree

Es un algoritmo de entrenamiento predictivo el cual permite dividir un gran conjunto de datos en un conjunto más sencillo mediante la aplicación de una secuencia de reglas de decisión en donde los nodos están representados por las condiciones o características del conjunto de datos, las ramas están relacionados con las decisiones que conducen a las etiquetas de la clase (si o no) y finalmente las hojas representan la etiqueta que se asignará una vez cumplida la condición como se muestra en la figura 2.5. Es un algoritmo de entrenamiento supervisado usado en su mayoría para tareas de regresión y clasificación como se indica en (Ullah, 2019).

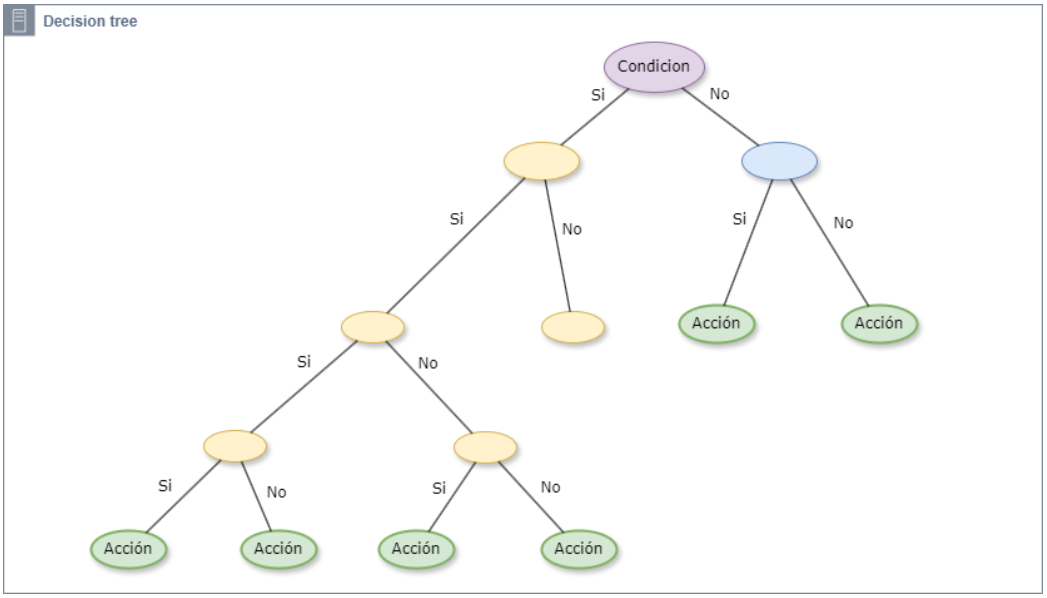

Figura 2.5:Funcionamiento de Decision tree.

Random Forest

Es un algoritmo de aprendizaje automático el cual está conformado por varios árboles de decisión de varios subconjuntos de datos con los cuales se genera un promedio o una votación mayoritaria de todas las predicciones obtenida de los árboles de decisión para mejorar el rendimiento del modelo obtenido con este conjunto de datos como se representa en la figura 2.6.

Se debe considerar que, a mayor número de árboles, mayor precisión obtiene el modelo en la capacidad de resolución de la problemática evitando de esta forma el overfitting. Este algoritmo es usado en su mayoría para tareas de regresión y clasificación lo cual se explicó en (Xie, 2009).

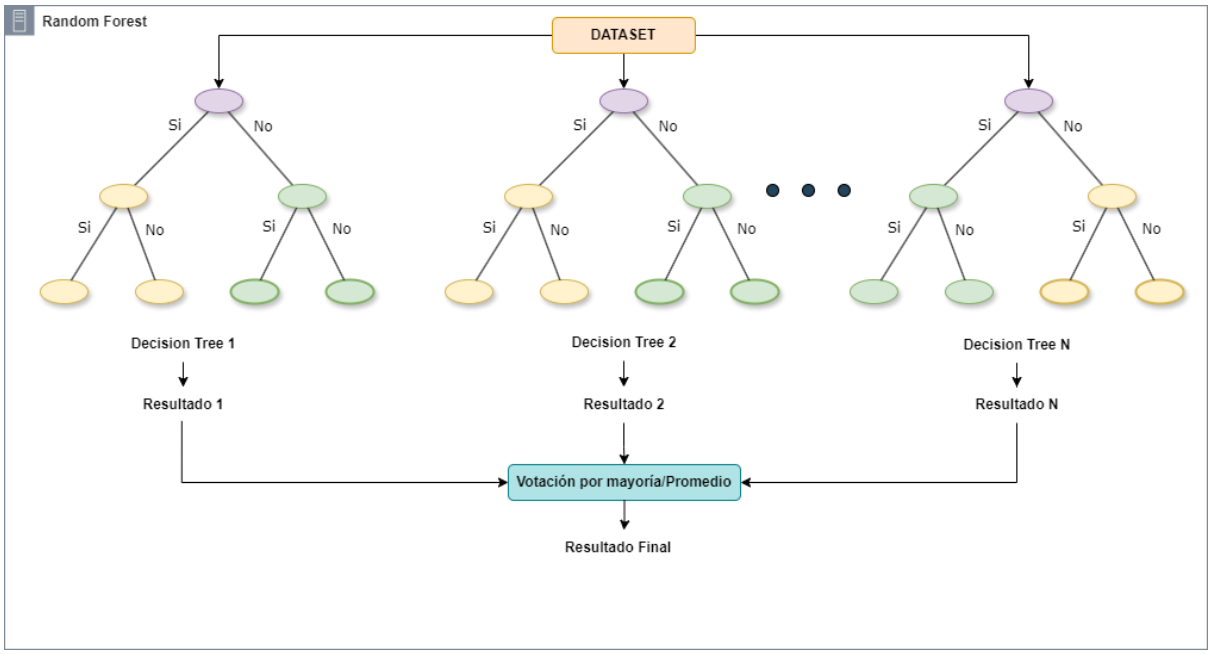

Figura 2.6:Funcionamiento de Random Forest.

Ensemble models

Los modelos de ensamble para el aprendizaje automático crean múltiples instancias de los algoritmos de aprendizaje automático más utilizados para generar una solución óptima a un problema con la capacidad de producir un modelo predictivo mejorado en

comparación con los algoritmos tradicionales obteniendo de esta forma un rendimiento superior como se presenta en la figura 2.7.

Los principales motivos para emplear modelos de ensamble es la incertidumbre en la representación de los datos o la existencia de una semilla inicial aleatoria en nuestro modelo.

Entre los modelos de ensamble más utilizados según (Pristyanto, 2022) tenemos los siguientes:

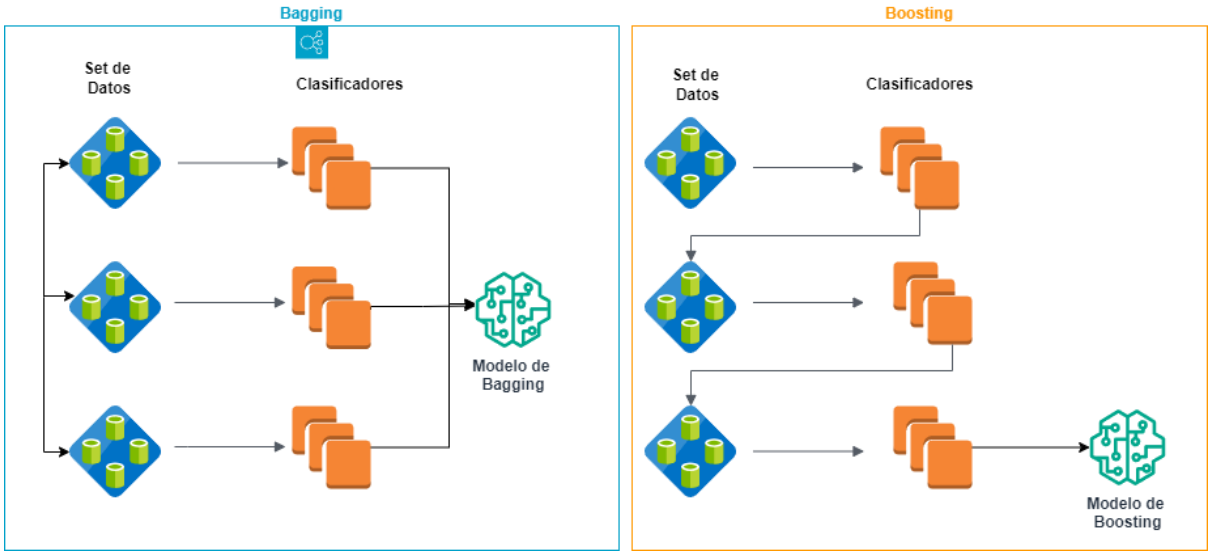

Figura 2.7:Funcionamiento de Ensemble models.

Bagging

El algoritmo de Bagging se basa en el uso de varios algoritmos simples de manera paralela en donde cada uno de estos representa un modelo simple e independiente con el objetivo de reducir el error al combinar sus resultados, lo cual conduce a una mejora en la precisión y estabilidad del modelo final (Yap, 2013).

Bagging emplea diferentes técnicas según el tipo de problema:

- En métodos de clasificación, se utiliza la votación mayoritaria, donde la clase final predicha es la que recibe más votos entre todos los modelos.
- En métodos de regresión, se utiliza el promedio de las salidas de los modelos, obteniendo así una estimación más precisa.

Boosting

El algoritmo de Boosting es una técnica de ensamble usada en aprendizaje automático para mejorar la precisión de los modelos (Imani M, 2023). A diferencia de Bagging que entrena modelos simples de manera independiente, Boosting lo realiza de manera secuencial comenzando con un modelo simple y luego ajustando los pesos de las observaciones y volviendolo a entrenar.

Esto se repite varias veces para realizar una media ponderada y de esta forma obtener el un modelo fuerte que mejora la precisión y rendimiento comparado con los modelos débiles individuales.

Stacking

Se determina como una técnica que mejora la capacidad predictiva de escenarios con data desbalanceada, la cual obtiene resultados eficientes combinando varios modelos de aprendizaje automático. Posee 4 principios fundamentales, los cuales son:

Ensamblado de modelos base: Se basa en el entrenamiento de varios modelos determinándolos como “modelos base”, en donde cada uno de estos aprende a realizar predicción de la data original, sin tener en cuenta el desbalanceo inicial de los datos.

Combinación de predicciones: realiza la combinación de todas las predicciones de los modelos base, uniendo las salidas de estos modelos.

Modelo meta: genera un nuevo modelo llamado “modelo meta”, a partir del aprendizaje obtenido de los errores de los modelos base. El modelo meta puede aprender a corregir estos sesgos combinando las predicciones y capturando patrones más complejos.
(Gore, 2023)

Manejo del desbalanceo: con la aplicación del modelo meta que realiza el aprendizaje a partir de la combinatoria de las predicciones de varios modelos y la captura de los patrones complejos, permite minorizar la dificultad del desbalanceo en las clases minoritas.

2.4 Librerias y software a utilizar

Python

Python es un lenguaje de programación muy popular por su sencillez y versatilidad. Se utiliza en muchas áreas, incluyendo la ciencia de datos, desarrollo web, automatización y más, debido a su sintaxis clara y una gran cantidad de librerías disponibles.

Matplotlib

Matplotlib es una librería de Python para crear gráficos y visualizaciones de datos. Es muy flexible y permite generar una amplia variedad de gráficos, desde simples gráficos de líneas y barras hasta complejas visualizaciones tridimensionales. Es una herramienta fundamental para el análisis y presentación de datos en ciencia de datos.

Seaborn

Seaborn es una librería de Python construida sobre Matplotlib que facilita la creación de gráficos estadísticos atractivos y fáciles de interpretar. Proporciona una interfaz de alto nivel para dibujar gráficos informativos y es especialmente útil para visualizar datos estadísticos de manera eficiente y estética.

Sklearn

Sklearn (o Scikit-learn) es una librería de Python que facilita la implementación de algoritmos de aprendizaje automático. Incluye herramientas para la clasificación, regresión, agrupamiento y reducción de la dimensión, y es muy utilizada en proyectos de ciencia de datos.

Imbalanced Learn

Imbalanced Learn es una librería de Python diseñada para tratar problemas de desequilibrio de clases en conjuntos de datos. Proporciona varias técnicas para el sobremuestreo, submuestreo y la generación de nuevas muestras, ayudando a mejorar la precisión de los modelos de aprendizaje automático en casos donde las clases están desproporcionadamente representadas.

Pandas

Pandas es una librería de Python que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento y fáciles de usar. Es especialmente conocida por sus DataFrames, que permiten manipular y analizar datos de manera eficiente.

Streamlit

Streamlit es una librería de Python que permite crear aplicaciones web interactivas y visualizaciones de datos de manera muy sencilla. Es muy utilizada en la ciencia de datos para construir y compartir rápidamente dashboards y aplicaciones que muestren análisis y modelos de datos.

CAPÍTULO 3

3 DISEÑO E IMPLEMENTACIÓN

3.1 Exploración y validación de datos y fuentes.

3.1.1 Dataset de ventas

Este conjunto de datos se obtuvo directamente del portal de ventas de la empresa en donde se detalla toda la información de las transacciones de compra de diferentes firmas existentes que seleccione el cliente.

Limitantes:

Dentro de este dataset se encuentra información desde el año 2022 al 2023 debido a que no se puede obtener registros más antiguos de las transacciones de clientes por los siguientes motivos:

- **Limitación de rendimiento de la base**: Las bases de datos del portal de ventas de los primeros años de la empresa no se encontraban optimizadas y preparadas para manejar un gran volumen de datos y consultas complejas, lo cual generó que el proceso de descarga de esta información sea lento y poco eficiente. Al ser bases de datos antiguas no eran compatibles con herramientas de extracción de datos actuales ya que ocasionaba conflictos en la migración de la información a sistemas más actuales, razón por la cual no se pudo obtener información de años anteriores al 2022.
- **Formato de dato obsoleto:** Debido a que los tipos de datos antiguos no coincidían con los actuales por los múltiples cambios en las estructuras de los datos a lo largo de los años lo cual genera una discrepancia en cuanto a las columnas de interés de nuestro estudio.
- **Explicación de los datos:** La ausencia de documentación clara y actualizada acerca de la estructura y funcionalidades de la base de datos puede dificultar su extracción y comprensión.

3.1.2 Dataset de las encuestas de satisfacción

Este dataset es obtenido a partir de un formulario de satisfacción de la venta del producto que recibe el cliente al generar la descarga de la firma, el llenado de esta información es obligatorio para todos los clientes que descargan la firma electrónica.

Limitantes:

- **Errores humanos en ingreso de datos**: En el ingreso de datos del formulario se pueden introducir datos erróneos seleccionando una opción incorrecta o proporcionando una información incompleta.
- **Duplicidad y nulidad de datos**: En los registros de los datos puede existir presencia de datos duplicados lo cual provoca confusión en la depuración de estos. De igual forma puede existir registros en donde el contenido de la mayoría de las columnas es nulo y no se puede obtener información relevante, en donde solución sería eliminar estas filas reduciendo de esta forma el número de registros de nuestro dataset.
- **Falta de estandarización:** La información suministrada por los clientes no siguen un formato estandarizado, principalmente en los campos de texto libre como las sugerencias para mejorar que se le consultan al cliente. La ausencia de estandarización complica la agregación y comparación de datos.
- **Sesgo en la captura de datos:** La estructura del formulario y la redacción de las preguntas y respuestas que puede seleccionar el cliente en el formulario genera sesgos que afectan las respuestas de los clientes ya que no reflejan con precisión la realidad o las opiniones de los clientes.

3.1.3 Dataset de soporte al cliente

Este dataset es obtenido a través de una herramienta de omnicanalidad en donde se detallan los soportes realizados a los clientes con el respectivo departamento asignado y el motivo de soporte.

Limitantes:

- **Estandarización de los datos:** La estructura de algunas columnas las cuales pueden tener un arreglo con los diferentes tipos de soportes, el cual va incrementando con cada soporte lo cual dificulta el conteo y preprocesamiento de la información.
- **Duplicidad de los datos:** Los datos de algunas columnas presentan variaciones que poseen el mismo significado lo cual limita la correcta clasificación de los datos.
- **Nulidad de los datos:** La presencia de filas enteras con datos nulos provoca la reducción significativa del tamaño total de nuestro dataset.

3.2 Analítica de datos

En la elaboración del aplicativo de predicción de abandono de clientes es de vital importancia la exploración de los datos transaccionales de los clientes, obteniendo como resultado de este análisis exploración, información relevante para nuestro estudio, que ayudará a un correcto filtrado de los datos a utilizar.
A continuación, se detallan algunos análisis realizados para una correcta segmentación de los datos:

En la figura 3.1 se detallan los resultados obtenidos en el análisis de volumen de venta de firmas electrónicas que se registran en las ventas de totalidad de los años mencionados (2022-2023), se puede evidenciar que las vigencias de la firma electrónica a partir 1 a 4 años presentan un mayor volumen de ventas, a diferencia de las vigencias de firmas de 1 mes, 6 meses y 5 años. Por este motivo se está tomando los datos con mayor volumen de ventas para el estudio.

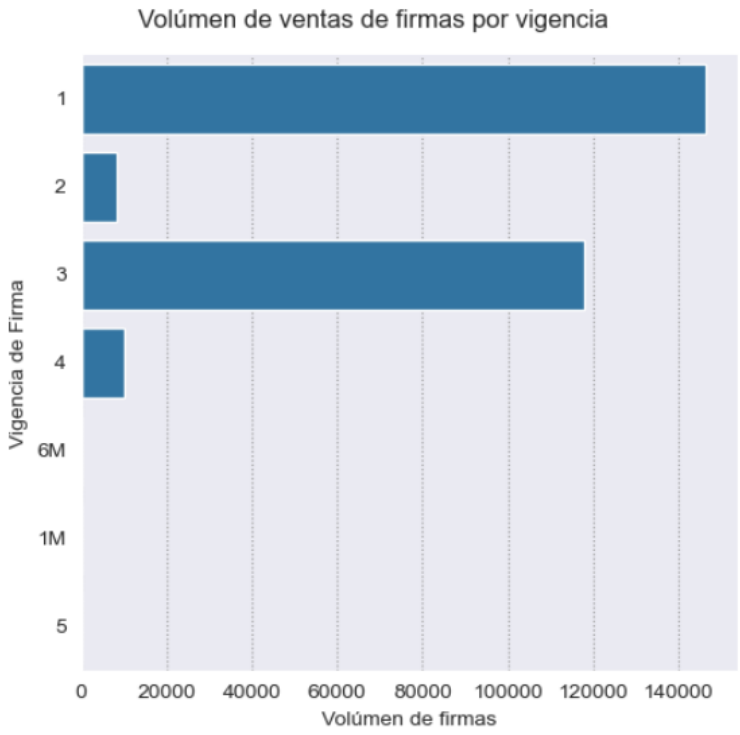

Figura 3.1:Ventas de firmas por vigencia.

En este análisis de la comparativa de clientes nuevos que se realizó con los datos obtenidos desde 2021 a 2023, se puede evidenciar en la figura 3.2 que existe una gran diferencia en el volumen de clientes nuevos que la empresa adquirió del año 2021 al 2022, debido al cambio de regulación de la facturación electrónica en el país, en donde se volvió obligatorio la facturación electrónica, para lo cual era indispensable una firma electrónica para la emisión de los comprobantes autorizados. Y en comparativa del año 2022 al 2023 el incremento de clientes se evidencio en una reducción, en la cual enfocamos principalmente el análisis de nuestro estudio en estos dos años para comparar el número de nuevos clientes del año 2022 frente al año 2023, y poder determinar una comparativa de la renovación o adquisición de nuevos clientes, por tales razones descartamos la información del año 2021 porque no existen registros completos de los clientes que puedan brindar mayor información en el análisis del estudio.

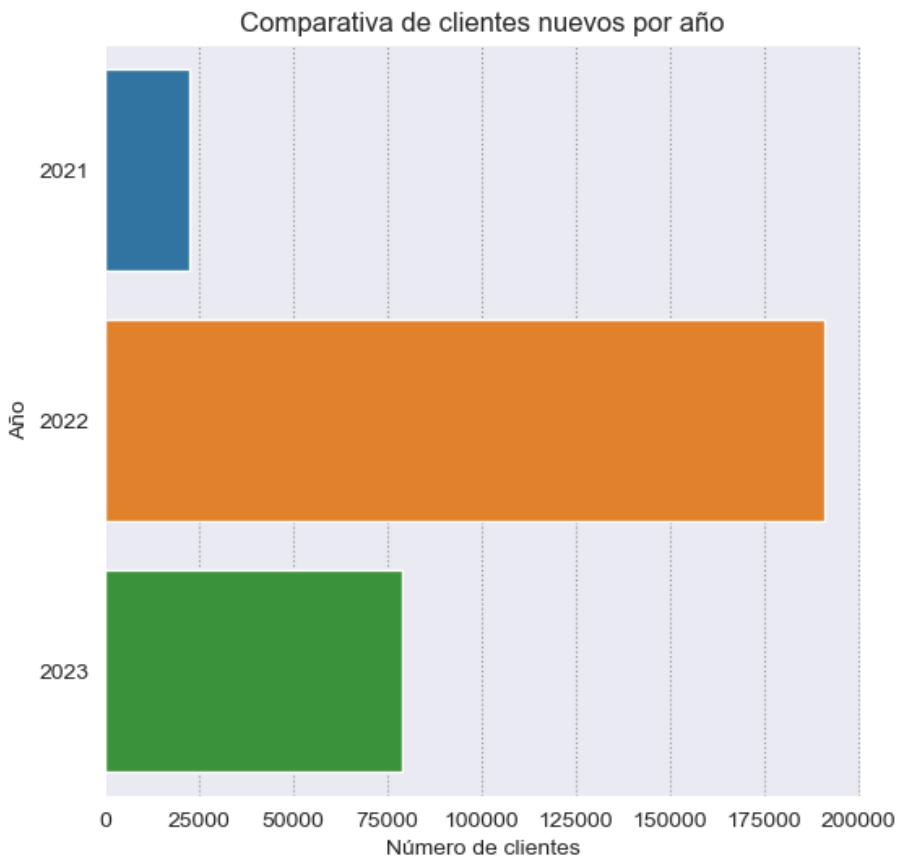

Figura 3.2:Clientes nuevos del 2021 al 2023.

En la figura 3.3 se establece un análisis de los tiempos de promedio de atención de los años 2022 y 2023 que se realiza para los diferentes tipos de servicios de atención (en línea y express) en los cuales está definido un promedio de atención por cada cliente, para otorgar un rendimiento alto de satisfacción al cliente en una rápida y efectiva atención.

Por ese motivo solo enfocamos el estudio en estos dos principales tipos de atenciones que maneja directamente la empresa, sin evaluar los otros tipos de atenciones que se genera por terceros vinculados, agentes o distribuidores de firmas electrónicas, los cuales poseen su propio tiempo de gestión de una solicitud de firma electrónica.

Como se puede evidenciar en el análisis de los promedios de atención, existe una variación entre estos principalmente en las solicitudes de las diferentes calidades de obtención de firma electrónica que es para personas naturales o jurídicas (representante legal o miembro de empresa), en la cual cada una de estas posee un proceso de revisión

diferente que implica un mayor número de pasos en el procesamiento de la solicitud para otorgar el servicio final.

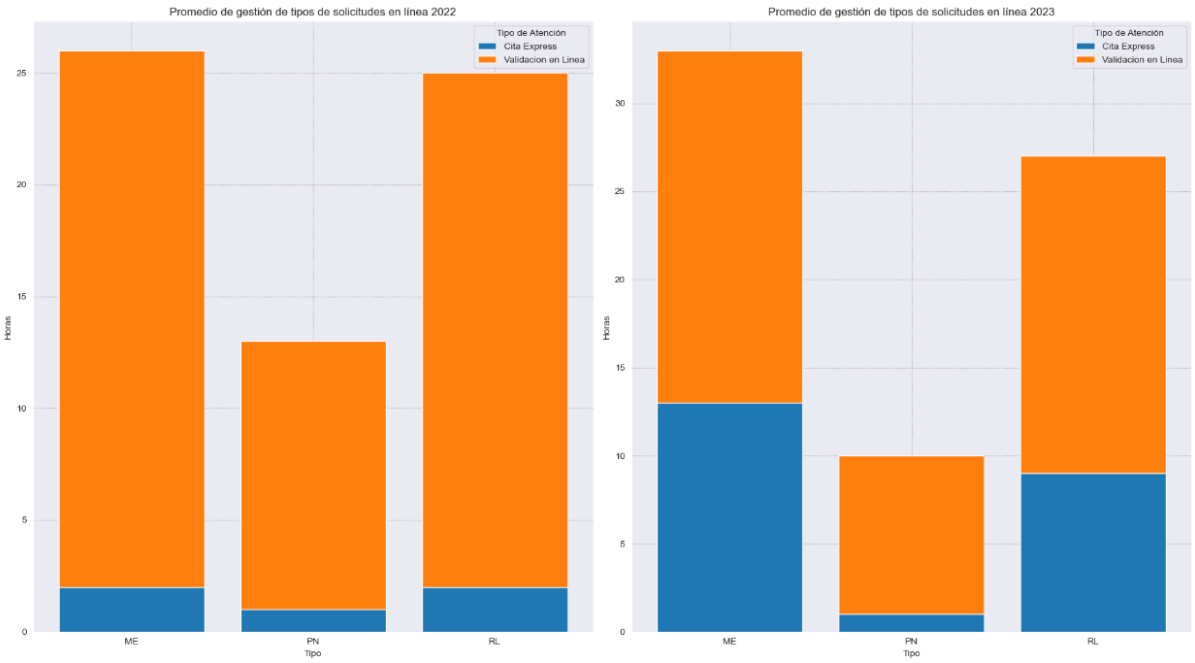

Figura 3.3:Promedios de gestión de atenciones en línea y express.

El Net Promoter Score (NPS) es una métrica utilizada para medir la lealtad y satisfacción de los clientes hacia una empresa. Se basa en una única pregunta: "En una escala de 0 a 10, ¿cuán probable es que recomiendes nuestro producto o servicio a un amigo o colega?" Las respuestas se clasifican en promotores (9-10), pasivos (7-8) y detractores (0-6). Esta métrica es beneficiosa para las empresas ya que proporciona una visión clara y directa de la satisfacción del cliente, identifica áreas de mejora, y permite establecer estrategias para aumentar la lealtad del cliente, lo que puede resultar en un crecimiento sostenido y mayores ingresos.

En el análisis realizado en la figura 3.4 se puede evidenciar que el score NPS del año 2022 es menor en comparativa de los otros años, y esto se debe por el aumento del volumen de venta de firmas en el año 2022, pudo haber generado una insatisfacción en el cliente por diferentes motivos tales como: tiempos altos de espera de resolución de soportes y de aprobaciones de solicitudes de firmas como se pudo observar en el grafico anterior del análisis promedio de tiempo de gestión de firmas.

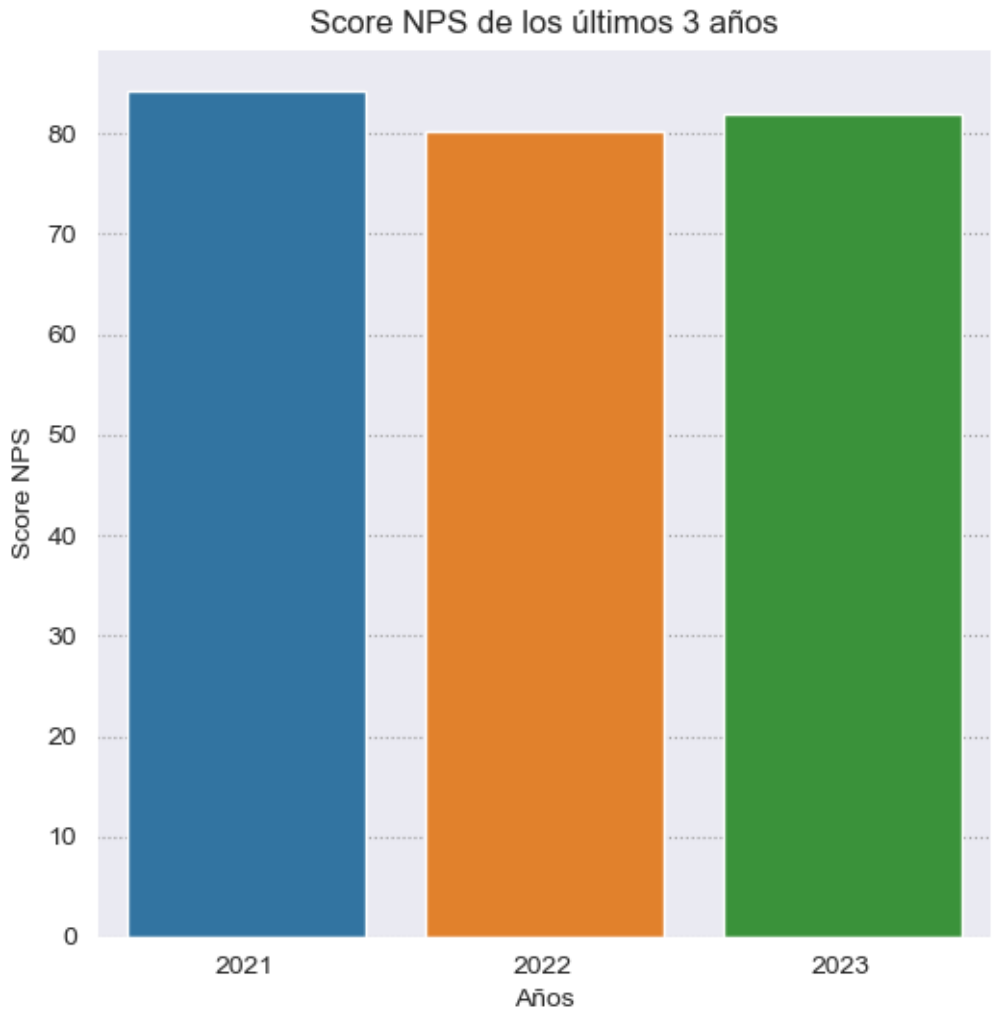

Figura 3.4:Score NPS de los 3 últimos años.

Se realizo un análisis de los soportes generados en el aplicativo de omnicanalidad de atención de clientes, evidenciando que los meses que existe un decrecimiento del número de soportes es de julio a agosto, pero a partir del mes de septiembre hasta el final del año la cantidad de soportes incrementa como se evidencia en la figura 3.5.

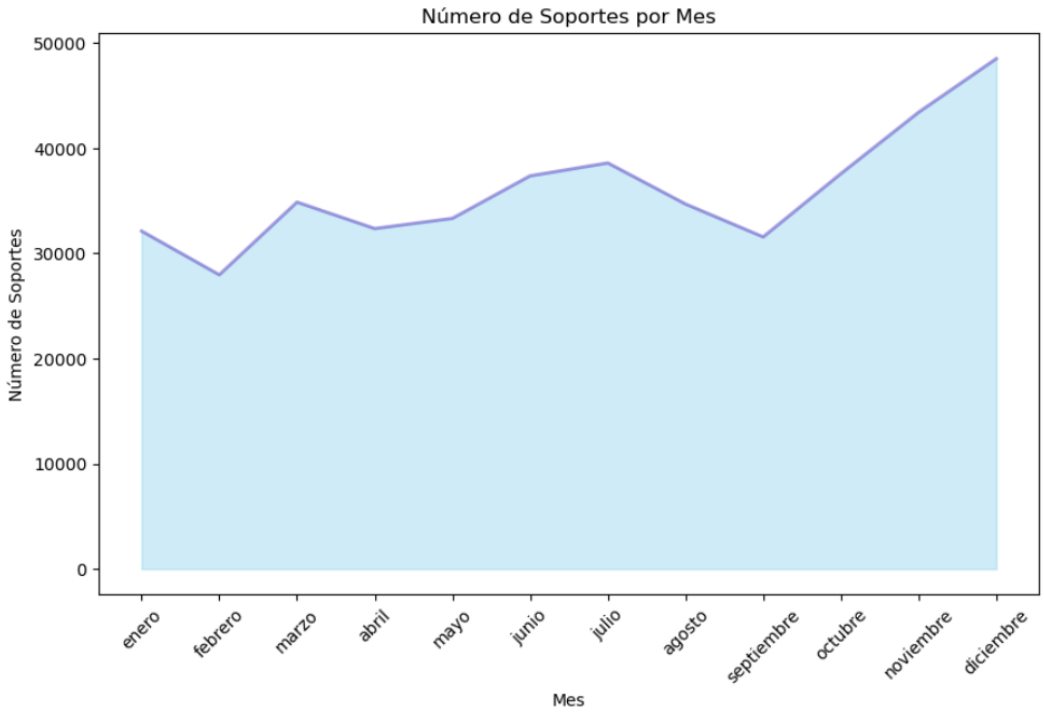

Figura 3.5:Total de soportes gestionados en 2023.

De tal forma como incrementa el número de soportes en estos meses, el tiempo promedio de atención de un cliente en cola para la revisión de su solicitud de soporte aumenta.

En el estudio del número de casos gestionados por los diferentes departamentos que generan soportes por medio de este aplicativo plasmado en la figura 3.6, se puede evidenciar que el departamento de renovaciones presenta un mayor número de gestiones de casos, debido a que utilizan este medio para gestionar la renovación del servicio del cliente, continuidad del proceso y finalización de la entrega del servicio.

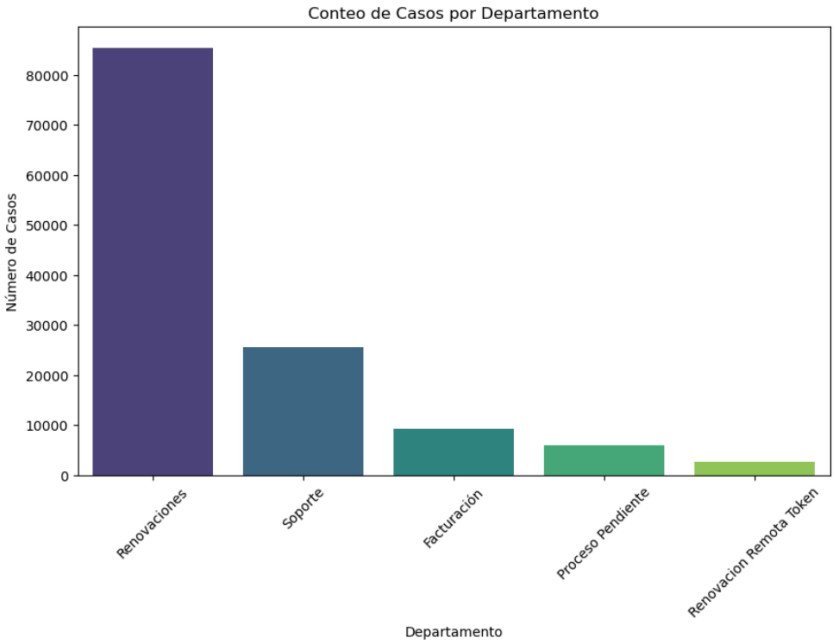

Figura 3.6:Casos gestionados por departamento.

Para el entrenamiento de nuestro modelo se puede evidenciar que el dataset final usado, presenta un claro desbalance entre las etiquetas de clientes perdidos y retenidos tal como se encuentra representado en la figura 3.7, para lo cual será necesario utilizar técnicas de balanceo de datos las cuales ya fueron detalladas con anterioridad en el capítulo 2.

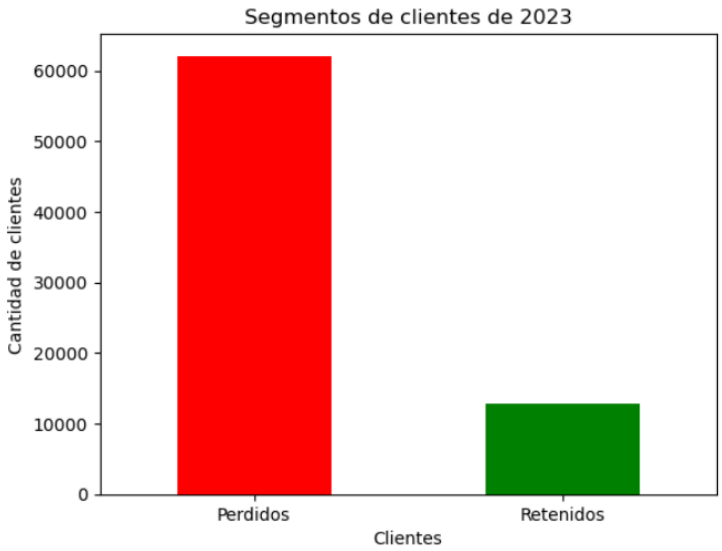

Figura 3.7:Segmentación de clientes perdidos y retenidos.

Se puede observar en la figura 3.8 el porcentaje de retraso en meses que tienen los clientes para generar una renovación del servicio, esto se validó verificando la fecha de caducidad de la firma electrónica y contrastándola con la fecha de renovación de los clientes.

En la gráfica del lado izquierdo se puede observar que el 49.1% de los clientes renuevan en el mes correspondiente de vigencia de la firma, llegando hasta un 90% aproximadamente en aquellos clientes que renuevan pasados 5 meses de su fecha de renovación.

En la gráfica del lado derecho podemos observar aquellos que renuevan mucho antes de la fecha de caducidad de su firma electrónica, lo cual puede tener distintas razones tales como: pérdida, olvido de contraseña, entre otros.

Por estas razones se decidió realizar la predicción de abandono de clientes tomando el año completo de vencimiento de las firmas electrónicas debido a que los clientes puede renovar hasta pasado 5 o 6 meses del vencimiento de su firma.

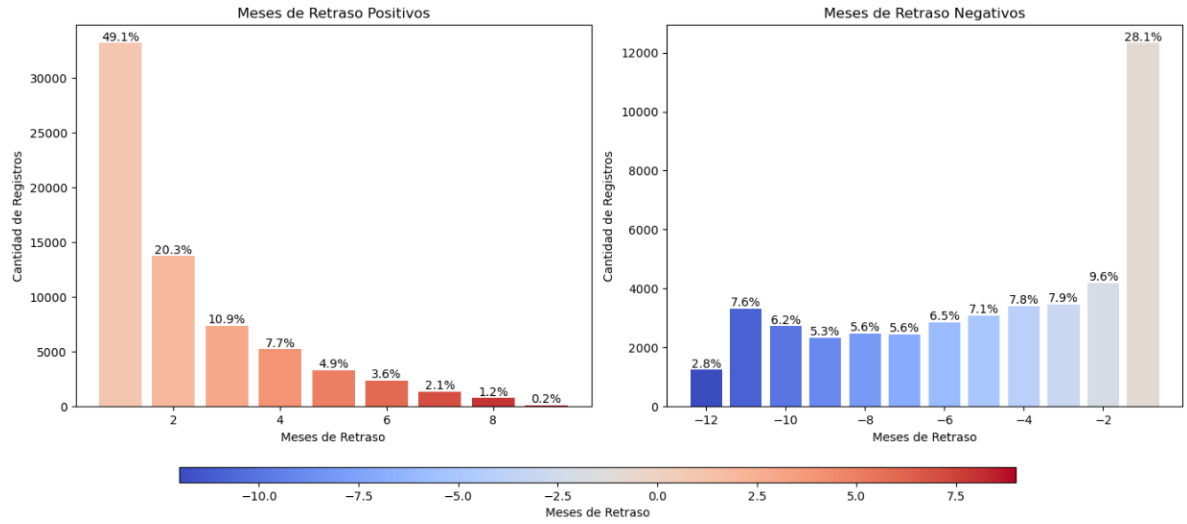

Figura 3.8:Porcentaje de retraso de renovación de firma.

3.3 Prototipos de algoritmos, modelos, y módulos del sistema

El entrenamiento del modelo de nuestro estudio está comprendido por diversas etapas esenciales que abarcan desde la comprensión del problema hasta la implementación final del modelo entrenado que nos permitan asegurar que el modelo sea preciso, robusto

y capaz de ofrecer insights valiosos para la retención de clientes. A continuación, se describen los pasos que se deben seguir.

1. Definición del problema del negocio: Involucrarse con las partes interesadas (Dpto. de operaciones) para identificar la importancia del abandono de clientes para la empresa y la afectación que tiene en los ingresos y satisfacción del cliente para finalmente predecir aquellos que posean una mayor probabilidad de abandonar el servicio.
2. Obtención de datos: Se recolecto datos de las fuentes directas de la empresa tales como lo de los portales de ventas del servicio, las encuestas de satisfacción de gestión del servicio al cliente y el dataset de los soportes que generan los asesores por el medio de omnicanalidad para aplicarlas en el estudio del modelo.
3. Exploración de los datos: Se realizo un estudio de los datos para comprender la distribución de los datos, mediante el uso de gráficos para la detección de las tendencias de incremente o decremento de las características primordiales que afecten en la tasa de abandono de clientes.
4. Aplicar procesos de transformación, depuración de datos entre los cuales se encuentran eliminar columnas y registros duplicados.
5. Consolidar las distintas fuentes de datos.
6. Realizar el preprocesamiento de datos: Conteo de registros de un tipo específico, crear nuevas columnas.
7. Crear los conjuntos de datos de entrenamiento y testeo.
8. Aplicar balanceo de clases utilizando SMOTE el cual elimina observaciones mal clasificadas y aquellas que tienen conflictos en sus vecinos más cercanos, mejorando la calidad del conjunto de datos balanceado.
9. Entrenar modelos de aprendizaje automático los cuales detallamos a continuación:
    - Decision tree
    - Random Forest
    - XGBoost
    - LightGBM
    - Catboost
    - Stacking

10. Optimizar los hiperparámetros de los modelos recopilados de la revisión de trabajos relacionados con la predicción de abandono de clientes los cuales se muestran en la tabla 3.1:

Modelos Hiperparámetros	Decision Tree	Random Forest	XGBoost	LightGBM	CatBoost
Max_depth	3,5,7	3,5,7	3,5,7	3,5,7	-
Min_samples_split	2,5,10				
N_estimators	-	50,100,200	50,100,200	50,100,200	-
Learning_rate	-	-	0.01,0.1,0.2	0.01,0.1,0.2	-
depth	-	-	-	-	3,5,7
iterations	-	-	-	-	100,200,500

Tabla 3.1:Hiperparámetros revisados en la literatura.

11. Entrenar un modelo de Stacking que junta los modelos para obtener un mejor rendimiento y verificar si este produce una mejora con respecto a los anteriores.
12. Escoger el mejor modelo basado en el rendimiento mostrado el cual se encuentra representado en la figura 3.9.

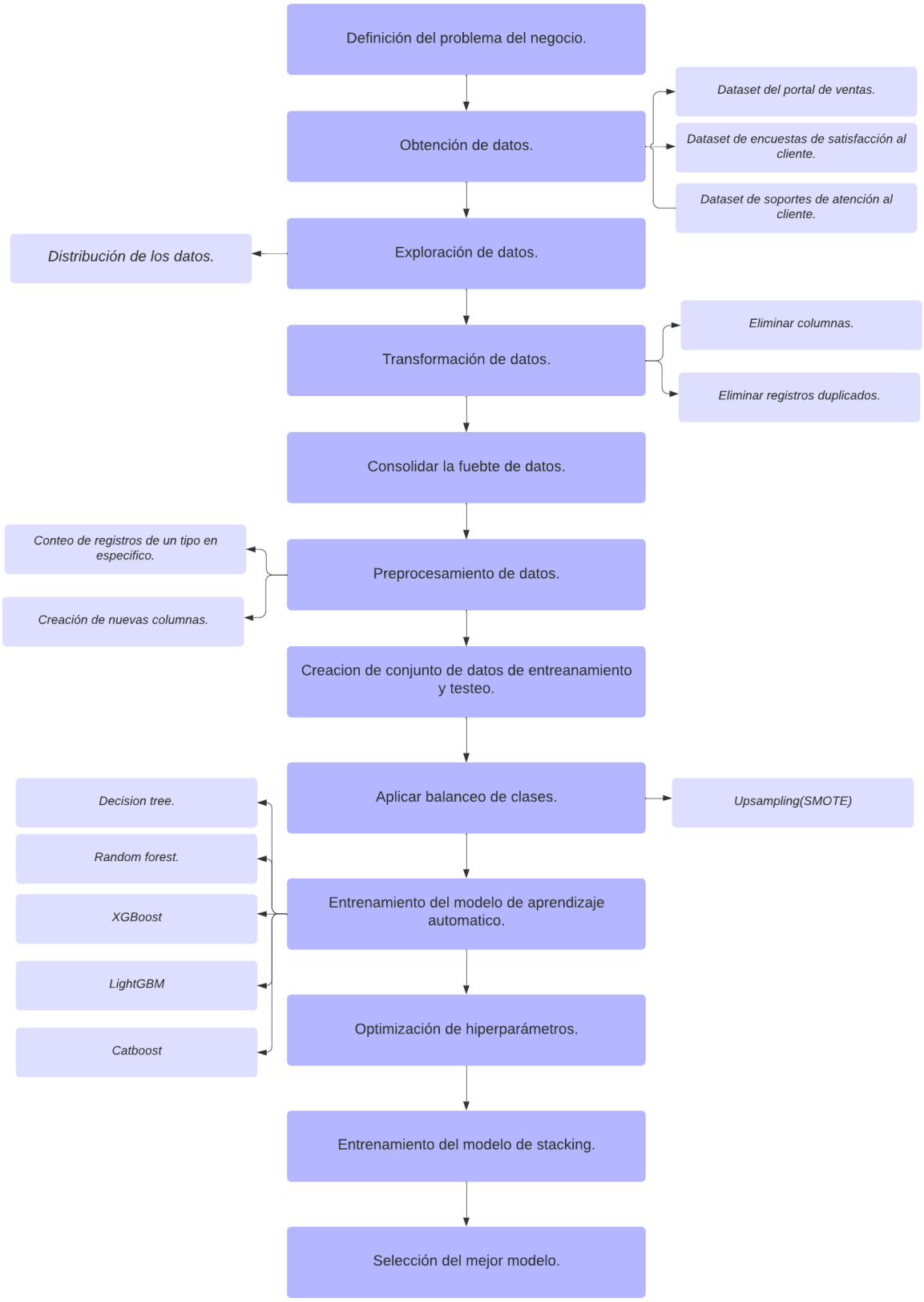

Figura 3.9:Diagrama de pasos de la creación del modelo.

3.4 Preprocesamiento de los datos

El preprocesamiento de datos es una etapa crucial para garantizar que la calidad de los datos analizados sea idónea para el análisis y modelado de nuestros datos permitiéndonos identificar las características relevantes de nuestro estudio, por tal razón hemos dividido el proceso en 4 etapas las cuales detallamos a continuación:

3.4.1 Depuración

Una parte esencial en el preprocesamiento es la limpieza de datos en donde se deben identificar y corregir las inconsistencias encontradas como duplicados o valores atípicos que puedan causar un problema en el análisis de nuestros datos. A continuación, detallamos la depuración realizada en cada uno de los dataset.

- **Dataset de portal de ventas:** En este dataset encontramos varias inconsistencias en cuanto a campos de fechas las cuales tenían formatos erróneos o incluso un texto cualquiera por lo que se procedió a la eliminación de las filas con valores faltantes o erróneos y en columnas repetidas se procedió a dejar una sola columna que fue la de creación de solicitud de firma.
  Otra de las novedades encontradas fueron los datos duplicados y valores atípicos en los valores de firmas electrónicas debido a la presencia de porcentajes de descuento o la venta masiva por lote de firmas lo cual no brindaba una proyección correcta de los precios brindados por la empresa por lo cual se procedió a realizar una imputación de valores usando la moda del tipo de firma adquirida para definir el valor de la factura del cliente.
  Como tercer punto se procedió a crear nuevas columnas a partir de otras ya existentes como obtener la provincia donde se obtuvo la firma a partir de la sucursal de venta.

- **Dataset de atención al cliente:** En este dataset encontramos varios valores faltantes en cuanto a la columna de comentarios puesto que este se trata de un campo opcional por lo que se procede a eliminar esta columna debido a que la

mayoría de los datos se encontraban vacíos y no representaba una información relevante para nuestro estudio.

- **Dataset de soportes al cliente:** En este dataset se encontraron algunos campos vacíos en la sección del tipo de soporte generado al cliente debido a que el proceso de asignación de una etiqueta para este soporte es realizado por el asesor y en algunos casos no se generaba una asignación. De este conjunto de datos únicamente se seleccionó los soportes generados a través de WhatsApp debido a que se generó una relación del número del cliente con el registrado en el portal encontrando mayores coincidencias de relación de soportes generados para cada cliente a diferencia de los otros canales de comunicación como: chat web, Instagram y Facebook en donde no se obtuvo información relevante para relacionar a los clientes.

Tras completar el preprocesamiento y etiquetado de los datos, se obtuvo el siguiente conjunto final, resultado de la integración de las tablas previamente mencionadas: encuestas de satisfacción, soportes y ventas de firmas. Durante este proceso, se eliminaron las columnas que aportaban poca información y se crearon nuevas, como el conteo de soportes realizados a cada cliente, para enriquecer el análisis.

Como se ilustra en la figura 3.10, la variable objetivo en nuestro estudio es la columna “Churn” la cual se obtiene por medio de la comparación de clientes del año 2022 que se encuentran en los registros del año 2023 o 2024 lo cual nos indica que si han renovado el servicio y que poseen una firma de 1 o 2 años. Se define a la variable “Churn” con el valor de 1 si el cliente no renueva la firma electrónica y como 0 cuando el cliente realizara la renovación de su firma.

Tipo	Pregunta 1	Pregunta 2	Pregunta 3	Renovación	Años	Nuevo Valor	Team	Etiquetas	Count	Churn
PN	Era Cliente	5	10	Si	1	17.92	Renovaciones	Facturado	3.0	0
PN	Era Cliente	5	9	No	1	19.04	Facturación	Consulta	2.0	0
PN	Recomendacion Servicio	5	10	No	1	19.04	Renovaciones	Información de Firma	7.0	1
PN	Redes Sociales	5	10	Si	1	17.92	Ninguno	Ninguno	0.0	1
RL	Era Cliente	5	10	No	1	19.04	Ninguno	Ninguno	0.0	0
RL	Era Cliente	5	10	Si	1	17.92	Renovaciones	Información de Firma	2.0	1
PN	Redes Sociales	4	9	No	1	19.04	Ninguno	Ninguno	0.0	1
PN	Redes Sociales	4	8	No	1	19.04	Ninguno	Ninguno	0.0	1
PN	Recomendacion Servicio	5	10	No	1	19.04	Ninguno	Ninguno	0.0	1
ME	Recomendacion Servicio	4	9	No	1	19.04	Ninguno	Ninguno	0.0	1

Figura 3.10:Tabla final para el entrenamiento del modelo.

3.4.2 One hot encoding

Una vez hecho el análisis y la depuración de los datos de nuestros datasets se procedió a realizar un merge de los datos para obtener un dataset unificado en donde se realizó la técnica de one hot enconding para la conversión de los datos categóricos en un formato numérico.

En la tabla 3.2 se muestra las características obtenidas del conjunto de datos:

Tipo_persona	Medio_conocimiento	Atención	Provincia	Equipo	Tipo_soporte
PN	Redes sociales	Express	Guayas	Facturación	Facturado
ME	Recomendación	En línea	Azuay	Renovación	Consulta
RL	Era cliente	express	Pichincha	Soporte	Adquirir firma

Tabla 3.2:Ejemplo de proceso de one hot encoding.

Luego del proceso de One Hot Encoding los datos de la tabla quedarían las dos primeras variables categóricas de nuestro conjunto de datos representados por las tablas 3.3 y 3.4:

Tipo_persona_PN	Tipo_persona_ME	Tipo_persona_RL
1	0	0
0	1	0
0	0	1

Tabla 3.3:Resultado de la variable tipo de persona.

Medio_Redes_Sociales	Medio_Recomendación	Medio_Era_Cliente
1	0	0
0	1	0
0	0	1

Tabla 3.4:Resultado de la variable de medio de conocimiento.

De esta manera se realiza el proceso de conversión de todas las columnas categóricas en valores numéricos aumentando de esta forma el número de columnas, pero facilitando el entrenamiento de nuestro modelo.

3.4.3 Estandarización

Se implemento la estandarización de nuestros datos para facilitar la interpretación de los resultados, y poder comparar la importancia de las características relevantes en nuestro estudio asegurando que las características contribuyan en el análisis. Se utilizó el StandardScaler de la librería scikit learn lo cual nos ayudó a estandarizar las características de nuestro conjunto de datos comprobando esta acción mediante la verificación de la media y la desviación estándar de cada una de nuestras variables numéricas.

3.4.4 Balanceo de datos

Una vez realizado todos estos procesos anteriores se evidenció el desbalance de las clases a predecir y mediante los trabajos relacionados en la sección del estado del arte los cuales recomendaban un balanceo de datos, se procedió a utilizar SMOTE el cual ha sido utilizado en (Gore, 2023) cuya información se detalla a continuación:

SMOTE genera muestras sintéticas de la clase minoritaria creando nuevos puntos entre los datos reales, lo que aumenta la cantidad de observaciones sin duplicarlas directamente como ocurre en el sobre-muestreo aleatorio.

3.5 Plataformas y prototipos de visualización

Una vez realizado el entrenamiento del modelo de aprendizaje automático que nos ofrece los mejores resultados posibles se procederá a la integración con Power Bi, una herramienta de visualización de datos en la cual se puede importar el modelo ya entrenado y aplicarlo al respectivo dataset para de esta forma poder mostrar las gráficas de interés.

3.5.1 Secciones del aplicativo

- **Filtro de tipo de persona:** Se podrá filtrar esta información dependiendo del tipo de persona ya sea natural, miembro de empresa o representante legal.

- **Valores de venta mensual:** Texto informativo referente la ganancia obtenida por las firmas vendidas en el mes finalizado.

- **Renovaciones por provincia:** Gráfico informativo en donde se visualiza la provincia con más renovaciones por mes.

- **Medios de información más utilizados:** Gráfico informativo que muestra el canal de comunicación más popular por el cual el cliente se entera de nuestro producto

- **Equipo con más soportes mensuales**: Gráfico explicativo en donde se muestra el equipo con más soportes gestionados en el mes.

- **Top 5 soportes más populares:** Gráfico donde se detalla los soportes más solicitados por los clientes.

3.5.2 Sección de predicciones

- **Valores de ganancia y pérdida:** Texto informativo que muestra la ganancia y pérdida dependiendo del abandono de clientes predicho.

- **Mejores y peores sucursales:** Gráfico explicativo de las sucursales donde se pierde la mayor cantidad de clientes y aquellas donde no se pierden clientes.

- **Top 5 temas de interés:** Muestra las etiquetas más solicitadas por las cuales se pueden perder clientes.

A continuación, en la figura 3.11, se presenta un prototipo de la sección general de nuestro aplicativo en donde se detallan algunas gráficas, posteriormente se dividirá en dos secciones para que sea más claro para el usuario la división entre la información actual y la predicha.

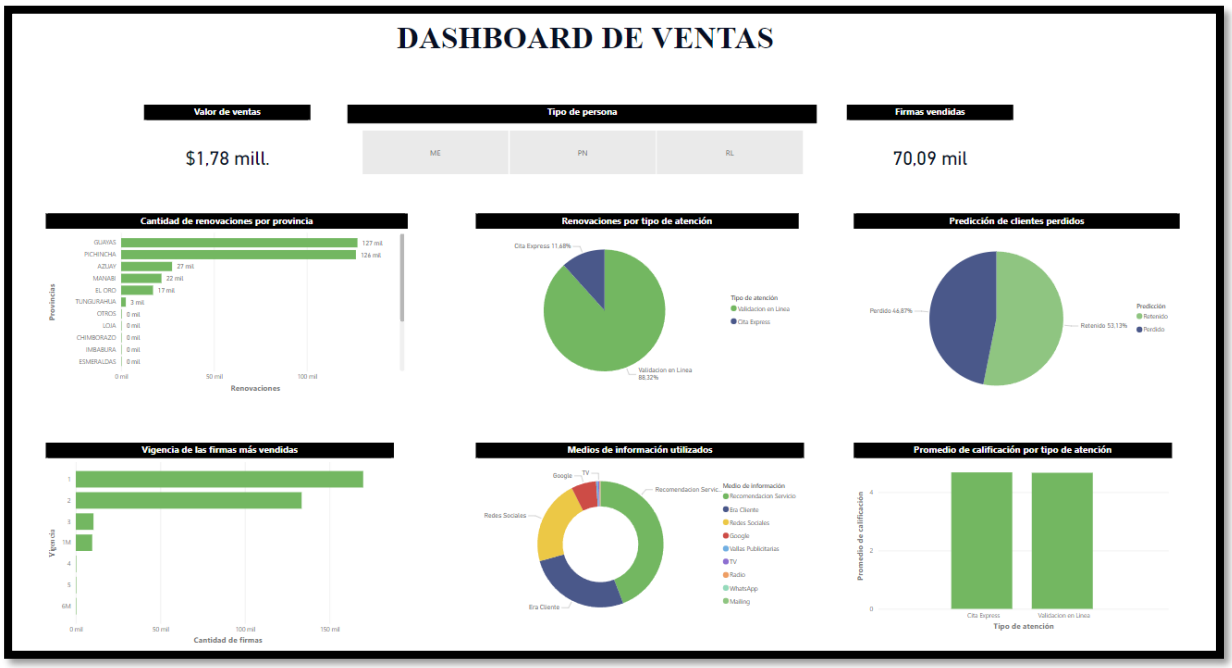

Figura 3.11:Prototipo de aplicativo.

3.6 Métricas y comunicación de resultados

Se realizo el análisis de la matriz de confusión, para la evaluación del rendimiento del modelo. En donde se clasifico como clase positiva a los clientes que abandonan y como clase negativa a los clientes que se quedan.

En la tabla 3.5 presentamos un ejemplo de la matriz de confusión orientada a nuestra problemática:

	Predicho: No Churn	Predicho: Churn
Real: No Churn	TN=300	FP=2000
Real: Churn	FN=60	TP=8000

Tabla 3.5:Prototipo de matriz de confusión.:Prototipo de matriz de confusión.

- **TP (verdadero positivo):** predijo correctamente que 8000 clientes abandonarían.
- **FN (falso negativo):** predijo incorrectamente que 60 clientes abandonarían.
- **TN (verdadero negativo):** predijo correctamente que 300 clientes no abandonarían.
- **FP (falso positivo):** predijo incorrectamente que 2000 clientes abandonarían.

Uno de los objetivos de nuestro análisis es obtener un alto TP y un bajo FN ya que esto nos va a permitir identificar correctamente los clientes que abandonan nuestro servicio.

3.6.1 F1 score

Una de las métricas que se implementó fue el F1 score para medir el análisis del rendimiento de nuestro modelo, especialmente para el desbalance de clases que se evidencia en nuestro estudio de abandono de clientes.

El F1 score combina dos métricas (precisión y recall) en donde la precisión nos indica el porcentaje de predicciones correctas y el recall el porcentaje de verdaderos positivos capturados por nuestro modelo.

Mediante la implementación de esta métrica se busca obtener un equilibrio entre la precisión y el recall y poder identificar correctamente la clase positiva minimizando los errores de clasificación del modelo.

3.6.2 Accuracy vs F1 score

El accuracy es la proporción del número de veces que un modelo acierta correctamente tanto en predicciones positivas como negativas sobre el número total de las predicciones realizadas enfocándose en indicar que tan bien funciona un modelo en términos generales.

Se puede identificar el accuracy como una de las métricas más utilizadas en problemas de clasificación

El accuracy puede ser engañoso debido a que es considerada como una métrica intuitiva en la mayor parte de la resolución de problemas de clasificación, pero para el caso de abandono de clientes la cual posee un desbalance de clases esta métrica no reflejaría adecuadamente la predicción o manejo de la clase minoritaria.

En cuanto a manejo de las clases de desbalanceadas es preferible enfocarse en el f1 score el cual me proporciona un mejor equilibrio entre las métricas de precisión y recall de ambas clases obteniendo de esta forma una evaluación más significativa del modelo resultante de predicción de abandono de cliente, en donde los falsos negativos pueden tener un costo bastante alto para la empresa y finalmente obtener una buena predicción de los verdaderos positivos que son los clientes que abandonan.

CAPÍTULO 4

4 ANÁLISIS DE RESULTADOS

En el presente capítulo se exponen los resultados obtenidos desde el análisis exploratorio de datos en el cual se presentaron diferentes limitaciones como los datos históricos y la complejidad en la interpretación de estos datos, para poder presentar los hallazgos claves de manera estructurada mediante la implementación de los métodos utilizados para la evaluación de los datos y finalmente obtener la precisión de modelos predictivos y la identificación de factores claves.

Y como resultado poder presentar un dashboard con un diseño intuitivo y que permita a las audiencias no técnicas comprender las visualizaciones de datos complejos de manera clara y accesible, facilitando la toma de decisiones rápidas dentro del departamento de operaciones.

4.1 Recolección de datos y estrategias para validación del proyecto

Como primer punto se empieza con un análisis exploratorio del dataset final utilizado para el dashboard, en el cual procederemos a describir los hallazgos encontrados y que son de principal interés del departamento de operaciones.

Como se observa en la figura 4.1, se realizó un análisis que identifica los casos de soporte más demandados durante el año, destacando especialmente aquellos relacionados con los departamentos de renovaciones y soporte, que son el eje principal de este estudio.

Se puede observar los diferentes casos gestionados a través del medio de comunicación de chat en el cual operan distintos equipos tales como renovaciones y soporte de firma.

Las etiquetas de: Facturado, culminación de proceso e información de firma son implementadas por el equipo de renovaciones para la gestión de los casos de WhatsApp marketing gestionadas para los clientes de renovaciones de firma.

Además de la etiqueta de adquirir firma que es implementada por el departamento de soporte en el cual se le indica al cliente el proceso de obtención de firma mediante los diferentes medios de obtención de firma que se ofrecen a los clientes.

Figura 4.1:Cantidad de soportes por etiquetas.

La figura 4.2 presenta el análisis del total de clientes retenidos o perdidos por provincia, destacándose que en Pichincha y Guayas se observa tanto un elevado nivel de fidelidad por parte de los clientes como un notable índice de deserción.

Observando estos detalles se puede implementar segmentación y personalización de estrategias, para la optimización de recursos y mejoras de la experiencia al cliente en estas provincias críticas. Y poder enfatizar al estudio de cuáles son esos factores que influyen en la lealtad del cliente, guiando decisiones estratégicas para el negocio.

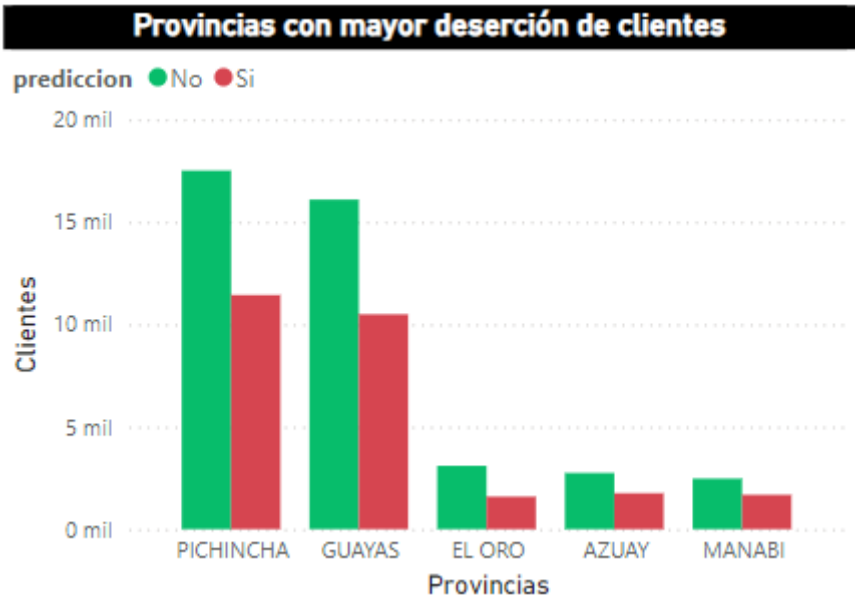

Figura 4.2:Provincias con mayor deserción de clientes.

Con la ayuda del siguiente gráfico podemos visualizar la concentración de segmentos de clientes que presentan mayor pérdida de clientes.

En el gráfico de barras de la figura 4.3 se hace referencia al análisis de deserción de clientes del servicio de firmado electrónico, en donde 47 mil clientes no han desertado y 29 mil sí lo han hecho, la deserción representa el 38.1% (29 mil de un total de 76 mil).

Este porcentaje obtenido en este estudio, indica una tasa significativa de pérdida de clientes que debe ser atendida con urgencia por la empresa con el objetivo de mejorar la retención de los clientes. Para poder abordar el análisis de los factores que podrían estar contribuyendo a esta deserción, como la calidad del servicio, precios altos, mejor oferta de la competencia o falta de personalización en la atención que recibe el cliente durante las etapas de venta y soporte.

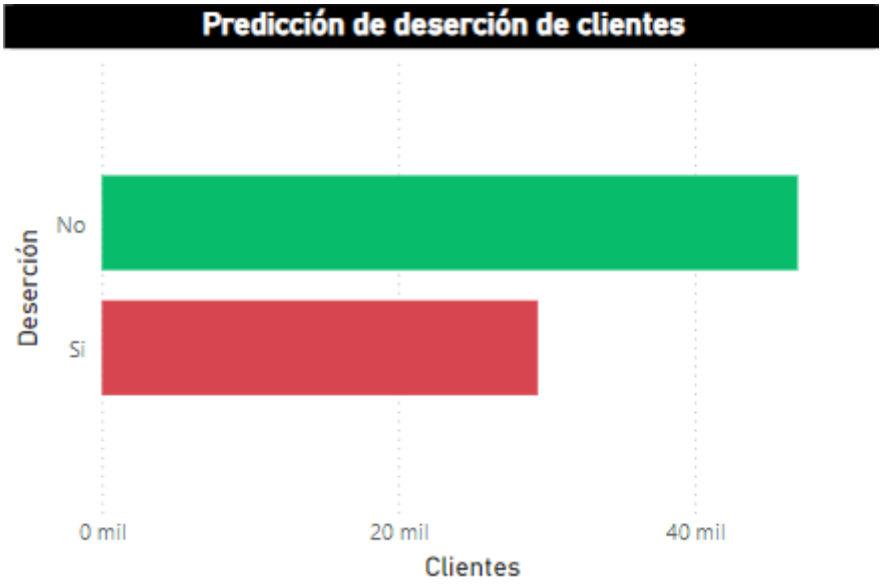

Figura 4.3: Deserción y retención de clientes.

Para poder evidenciar la cantidad de soportes o consultas que generan los clientes por los diferentes medios de comunicación, se realizó el análisis de la cantidad de soportes generados por mes durante el año 2023 mostrado en la figura 4.4, en donde se puede observar aquellos meses con mayor afluencia de personas que requieren de un soporte, se evidencia principalmente un incremento desde el mes de septiembre hasta diciembre, en donde los soportes incrementaron debido a la causa de adquisición o renovación del servicio de firma por la facturación electrónica que está despegando en el país, lo cual podremos analizar más adelante en el análisis de los tipos de soportes más generados por los clientes durante este periodo de tiempo.

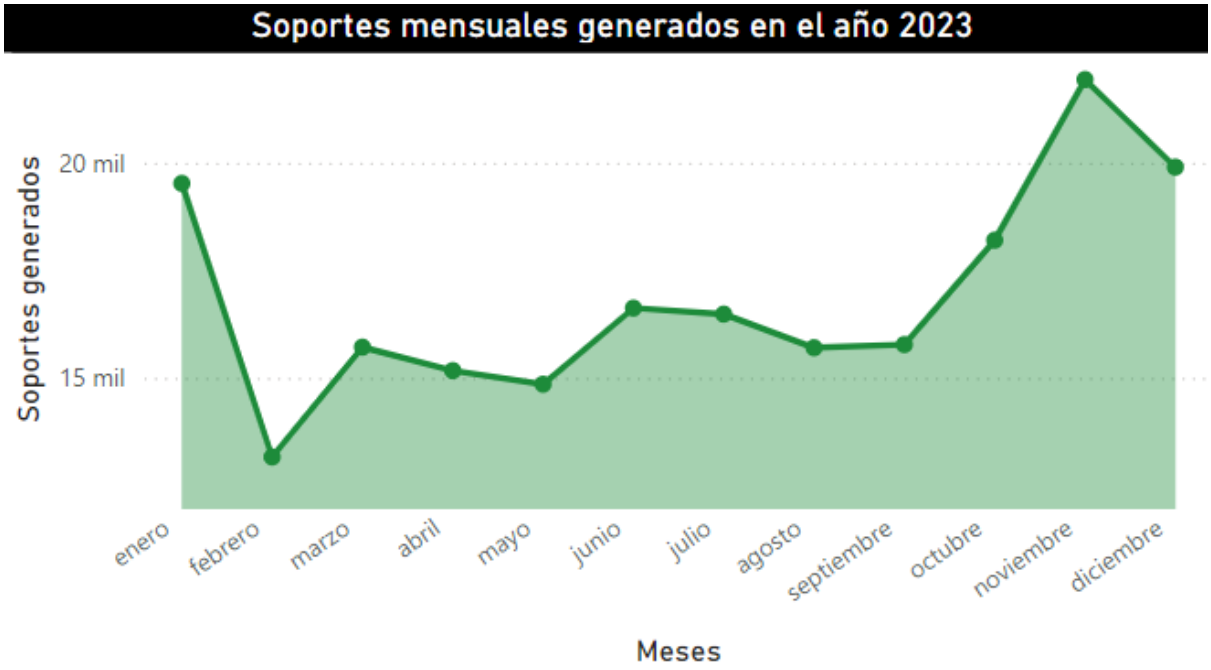

Figura 4.4:Soportes mensuales de 2023.

4.2 Resultados Finales del modelo

En esta sección, se presentan los resultados obtenidos tras entrenar varios modelos sobre el conjunto de datos final, utilizando técnicas como el one hot encoding, standard scaler y Smote, además de aplicar el balanceo de clases a cada modelo para evitar el overfitting.

La tabla 4.1 se resume los principales indicadores de rendimiento de cada modelo, incluyendo métricas como accuracy,precisión, recall, F1-score y AUC-ROC las cuales son aplicadas directamente en el dataset de validación las cuales son separados mediante muestreo aleatorio. Estos resultados nos permitirán evaluar el desempeño de cada modelo y seleccionar el más adecuado para la problemática de estudio, con base en su capacidad para generalizar sobre los datos de prueba y su ajuste a los objetivos del proyecto.

Modelo	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	0.557380	0.399719	0.757325	0.523259	0.653734
Random Forest	0.599430	0.416451	0.620302	0.498336	0.647546
XG Boost	0.420313	0.356027	0.998224	0.524858	0.654006
LightGBM	0.517513	0.386035	0.854099	0.531737	0.655142
CatBoost	0.421642	0.356341	0.996153	0.524912	0.657248
Bagging	0.636830	0.433620	0.432080	0.432849	0.632617
Stacking	0.639867	0.439838	0.448949	0.444347	0.622292

Tabla 4.1:Desempeño de los modelos.

En la figura 4.5 evidenciamos la matriz de confusión del modelo LightGBM ,el cual destaca como un modelo ganador en la predicción de abandono de clientes (churn) gracias a su capacidad para manejar datos grandes, desbalanceados y con características complejas, todo con alta precisión y velocidad. Este modelo se adapta especialmente bien a escenarios donde los datos incluyen múltiples variables categóricas y numéricas, ya que permite procesar directamente las características categóricas sin necesidad de codificaciones adicionales, lo que reduce la complejidad del preprocesamiento.

Además, su capacidad para ajustar parámetros como scale_pos_weight y is_unbalance lo hace efectivo en problemas de churn, que a menudo implican clases desbalanceadas. La combinación de su rapidez, eficiencia de memoria y alta interpretabilidad lo convierten en una herramienta ideal para identificar patrones complejos de comportamiento que predicen la retención o deserción de clientes, maximizando el impacto de las estrategias comerciales basadas en estas predicciones.

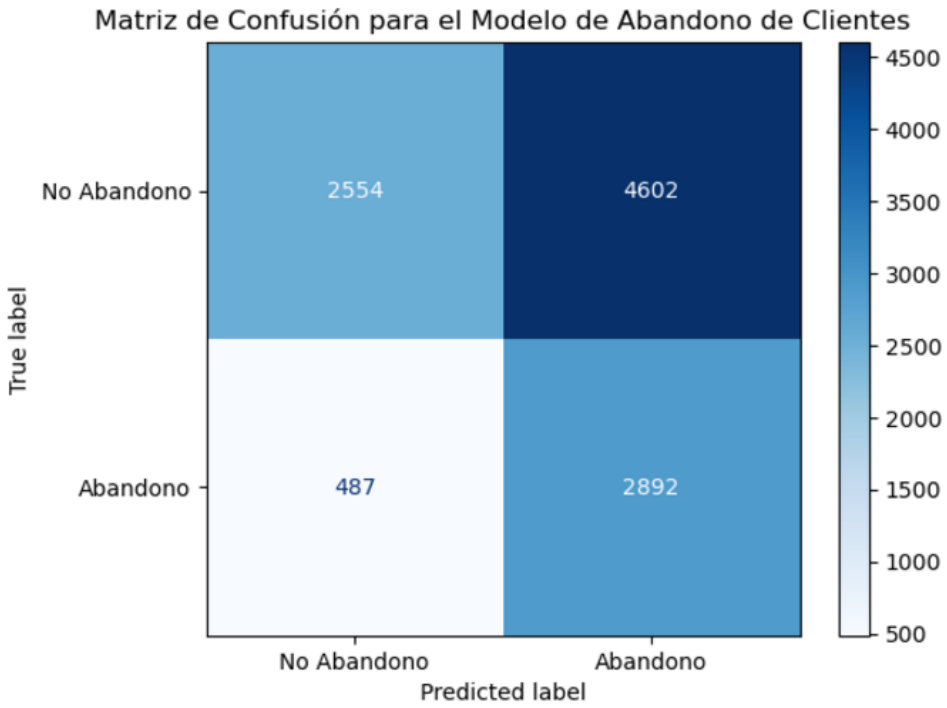

Figura 4.5:Matriz de confusión del modelo de LightGBM.

4.3 Importancia de variables

El gráfico de importancia de variables revela que la pregunta 1 que indica el método del conocimiento del producto el cual es el “era cliente” con un valor de 200. Esto indica que estás dando continuidad al servicio, lo que representa una retención exitosa por parte de la empresa. Esto refleja que el cliente no abandonó, lo cual es clave para la estrategia de fidelización.

El tipo de persona que aplica al proceso de renovación de firma en este caso esta el de RL que engloba a la renovación de solicitudes de representantes legales de empresa, la cual refleja un valor de 100. Esto indica que las estrategias de facilidad de renovación brindada a este grupo en específico tienen buenos resultados.

La atención brindada al cliente con el número de soportes brindados es uno de los factores más influyente en el modelo, con un valor cercano a 80, lo que indica su impacto significativo en la predicción del abandono de clientes. Este hallazgo sugiere que una atención de alta calidad es crucial para mantener la lealtad del cliente.

Finalmente, en la figura 4.6, se evidencia que estos resultados proporcionan una visión clara de los factores que afectan la retención de clientes, subrayando la importancia de una atención al cliente excepcional y relaciones sólidas.

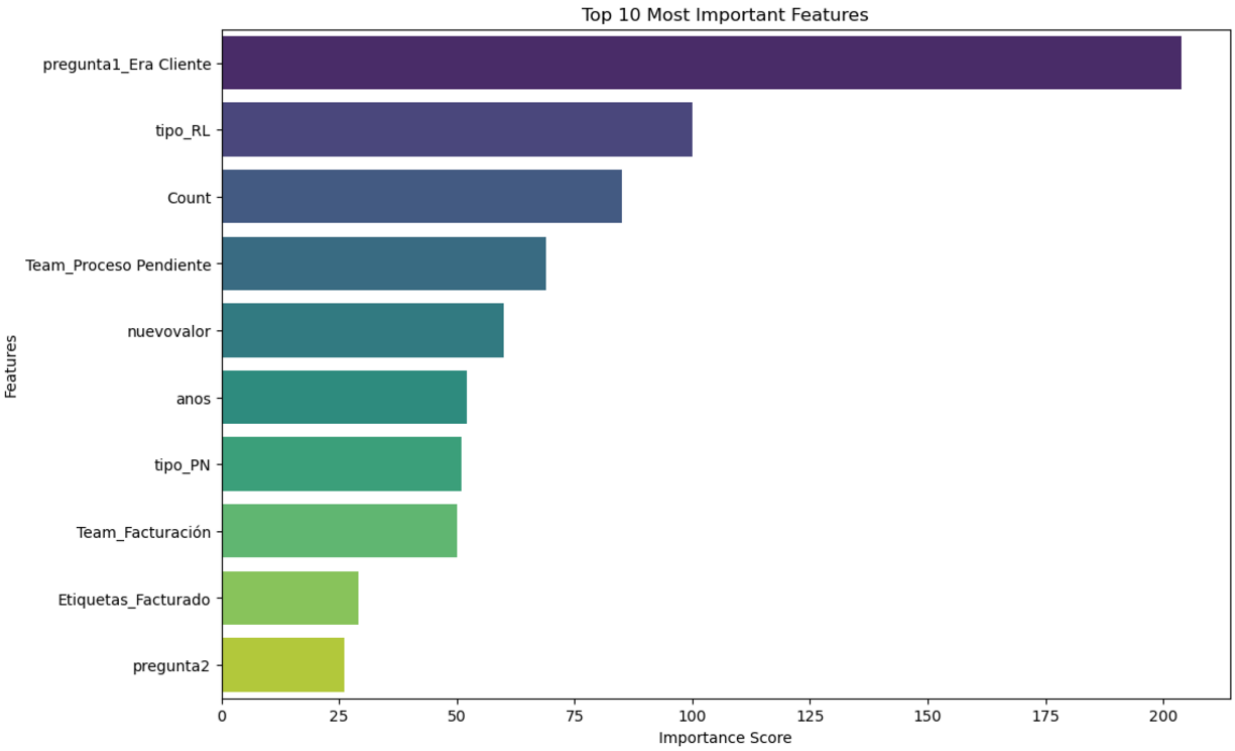

Figura 4.6:Importancia de las variables del modelo.

4.4 Puesta en marcha y funcionamiento

Una vez realizado el dataset final se procedió a realizar un dashboard de visualización en PowerBI en donde se visualizan tanto los aspectos descriptivos de la venta de firmas electrónicas en el año 2023, además del impacto de la deserción de clientes usando el modelo obtenido.

Este dataset se obtiene mediante la ejecución de todo el preprocesamiento de las fuentes de datos mencionadas en el estado del arte utilizando un script de Python directamente

en PowerBI lo cual no da como resultado el dataset que se utiliza en el dashboard junto con la respectiva predicción de abandono de clientes.

El análisis del dashboard representada en la figura 4.7, revela información clave sobre el comportamiento de renovación de los clientes. De un total de 76,000 clientes, 51,000 han decidido renovar sus servicios, lo que representa aproximadamente el **67.1%** del total. Por otro lado, 25,000 clientes han optado por no renovar, lo que equivale al **32.9%**. Estos porcentajes reflejan una tendencia positiva en la retención de clientes. Es crucial seguir monitoreando estas métricas para identificar áreas de mejora y continuar fortaleciendo la lealtad de los clientes.

Figura 4.7:Prototipo del dashboard.

4.5 Pruebas de funcionalidad

Una vez obtenido nuestro modelo funcional usando los datos desde 2022 a 2024 y la representación visual de la cantidad de clientes propensos a desertar se utilizó la técnica de BackTesting tomando como referencia el año 2022, específicamente los meses de

noviembre y diciembre para finalmente comparar los resultados en cuanto a renovaciones con nuestra solución.

En la tabla 4.2 que se muestra a continuación, se explica el análisis respectivo del churn de los clientes en los meses con mayor demanda del servicio. Cabe destacar que la variable de estudio “Churn” en el modelo se define en la validación de nuestro modelo con una probabilidad mayor a 0.48 en donde los falsos negativos tienen un mayor impacto que los falsos positivos, para aquellos clientes que abandonan en donde la métrica de clasificación seleccionada en nuestro estudio es el F1 score (representa el balance entre precisión y recall) la cual se ajusta al análisis de nuestro modelo ya que contiene un conjunto de datos desbalanceados.

Mes	Clientes totales	Clientes que abandonan	Clientes que abandonan (Modelo ML)
Noviembre	8954	3733	3169
Diciembre	8168	4468	2990

Tabla 4.2:Comparativas de abandonos identificados.

El modelo actual alcanza un F1 score de 0.47 en noviembre y 0.56 en diciembre, lo cual refleja una precisión moderada en la clasificación de clientes que retienen o abandonan. Este rendimiento puede estar influenciado por un desbalance de clases, la representatividad de los datos o los hiperparámetros seleccionados en el modelo base. Para mejorar estos resultados, proponemos una serie de ajustes: la optimización exhaustiva de hiperparámetros, la incorporación de nuevas características.

Para analizar el número total de clientes de estos meses se solicitó al departamento de renovaciones una base de datos que ellos manejan para realizar un segundo contacto con el cliente de los cuales no reciben una primera respuesta por medio de campañas de marketing por WhatsApp en el cual se genera un contacto directo con un asesor para ejecutar la renovación inmediata del servicio del cliente.

Como se puede observar en el mes de noviembre y diciembre de los clientes que abandonan, 3733 expresaron en noviembre que no deseaban el servicio y solo 4468 en diciembre que no deseaban continuar con la renovación.

Al implementar un modelo automatizado que aplica un modelo de Aprendizaje automático se puede identificar la mayor parte de la cantidad de clientes que abandonarán el servicio para aplicar un enfoque de recuperación de este cliente y reducir los tiempos de operatividad en realizar un contacto con el cliente.

4.6 Análisis costo/beneficio

De la base de datos de clientes que poseen de cada mes, envían un anuncio de marketing al cliente con un flujo enlazado al contacto de renovación del servicio con un agente, en el segundo recordatorio envían nuevamente al cliente la opción de contactar a un asesor para la renovación del servicio debido a que en el enlazado de acceso de estos mensajes tienen una duración de 72 horas tomando en cuenta que descartan aquellos clientes que si generaron una primera respuesta en el envío de la primera campaña.

Una vez contactados estos clientes para el ingreso de su solicitud de renovación, se genera una de cobranza para solicitar a estos clientes adjuntar sus pagos para finalizar el proceso de adquisición de la firma.

Se debe de tomar en cuenta que el resto de los mensajes que se pueden originar en la contabilización total de mensajes enviados por medio del aplicativo de campañas involucra a promociones de marketing u ofertas de servicios que hace el equipo de marketing el cual no es un punto de análisis en nuestro estudio.

Tal y como se evidencia en la tabla 4.3, se analizó el panorama actual en donde aproximadamente el 17% del envío de campañas representan clientes que abandonan el servicio de firma electrónica, debido a este antecedente se genera una exploración de

los planes disponibles de WhatsApp marketing con el objetivo de identificar un plan óptimo para los requerimientos del negocio involucrando el uso de nuestro modelo.

Por este motivo se propuso un nuevo plan de 35 mil mensajes de WhatsApp marketing lo cual tiene un costo aproximado de $2955.21 en comparación con el paquete de 52 mil mensajes lo cual representa un valor de $4460.20 lo que nos representa un ahorro del 33.74% en costos de envío de mensajes de WhatsApp marketing.

Mes	Total mensajes	Costo Paquete 52 K	Adicionales	Total
Noviembre	**52692**	**4391**	**69.20**	**4460.20**
Diciembre	**81537**	**4391**	**2953.70**	**7344.70**

Tabla 4.3:Costo actual de envío de campañas.

El cobro de costos por paquete de sesiones de campañas de marketing se define por un límite de mensajes generados los cuales se engloban a un solo valor fijo de 0.07 centavos dentro del paquete como se puede validar en la tabla un paquete de 52 mil mensajes tiene un valor total de $4391, pero el costo por mensajes adicionales que no entran en un paquete de sesiones se cobra 0.10 centavos por mensaje.

Se realizó un análisis costo/beneficio de la implementación de nuestra solución la cual toma en cuenta las siguientes estimaciones:

4.6.1 Beneficios

- Se midió el ahorro generado al enviar campañas de márketing para la renovación de clientes ya que el costo de envío por campaña depende de la capacidad de paquetes que se generen.
- Antes se enviaba una sola campaña de un segmento de clientes bien grande considerando dos recordatorios a los clientes para generar un primer contacto con ellos para la renovación del servicio.

- Si pasamos de realizar dos recordatorios de renovación de clientes, a una campaña de marketing de los clientes detectados que van a abandonar el servicio que sería un paquete de unos 5k y otro paquete semejante para el recordatorio debido a que con un primer contacto con promociones facilitaría la renovación de cierto segmento de clientes, lo cual implica un costo menor en envío de campañas de $2533 por campaña de renovación y $422.21 por campaña de recuperación de clientes, lo cual da un total de $2955.21 lo cual mostramos en la tabla 4.4:

Paquete	Costo
5K+30K	2955.21
52K	4460.20
Ahorro	**1504.99**

Tabla 4.4:Comparación de costos y beneficio obtenido.

Existen otros beneficios que no generan un valor monetario directo en el corto plazo, pero impactan indirectamente en factores que aumentan ingresos o reducen costos con el tiempo y que vale la pena mencionar tales como:

- Aumento en el indicador NPS de la empresa al retener una mayor cantidad de clientes.
- Aumento de recomendaciones y por ende aumento en nuevos clientes por una mejora en la calidad de atención.

4.6.2 Costos:

Recursos informáticos: $0

Debido a que la empresa posee una buena infraestructura tecnológica y cuenta con servidores y licencias activas de PowerBI no se lo incluye como costo en el proyecto

Recursos humanos:

Científico de datos: $1600 durante 2 meses para un total de $3200.

4.6.3 Tabla de costos/beneficios y retorno de la inversión

Se puede observar en la tabla 4.5 que se obtiene un retorno de la inversión a partir del tercer mes con los costos y beneficios mencionados anteriormente en donde se destaca el beneficio de ahorro en costos de campañas y el costo de recursos humanos para el proyecto.

	1er mes	2 do mes	3er mes	4to mes	5to mes
Beneficio Acumulado	1504.99	1504.99	1504.99	1504.99+504.97	1504.99+1739.96
Costo Acumulado	1870	1870+365.01	270+730.02	270	270
Diferencia	-365.01	-730.02	504.97	1739.96	2974.95
Retorno	No	No	Si	Si	Si

Tabla 4.5:Tabla de retorno de la inversión.

4.6.4 Análisis de costo por heurística actual de la empresa vs el modelo predictivo

La empresa actualmente no cuenta con un método eficiente para detectar el abandono de clientes. Por lo tanto, al implementar una campaña de renovación, envía mensajes masivos de WhatsApp a todos los clientes cuya fecha de caducidad del servicio está próxima.

Dado que no se dispone de un mecanismo para segmentar los mensajes y dirigirlos exclusivamente a clientes con mayor riesgo de abandono, se propone una heurística básica para evaluar el escenario actual. Esta heurística considera el costo asociado al envío masivo de mensajes por usuario, así como el valor recuperado por cada cliente que decide renovar el servicio gracias a esta campaña.

Para determinar la matriz de confusión de la heurística que envía mensajes a todos los clientes que les toca renovar, necesitamos calcular las métricas asumiendo que el modelo predice siempre que el cliente va a renovar (positivo). Esto significa:

1. **Verdaderos positivos (TP):** Todos los clientes que realmente renovaron (clase positiva).
2. **Falsos positivos (FP):** Todos los clientes que no renovaron, pero el modelo predice que sí.
3. **Falsos negativos (FN):** Ninguno, porque el modelo predice siempre "sí".
4. **Verdaderos negativos (TN):** Ninguno, porque no se predice "no" nunca.

Cada mensaje tiene un costo de 0.10 dólares por usuario.

Dado esto, la matriz de confusión de la heurística sería:

$$\begin{Bmatrix} \mathbf{4602} & \mathbf{0} \\ \mathbf{3379} & \mathbf{0} \end{Bmatrix}$$

- **Donde:**3379=487+2892
- **3379=487+2892:** todos los clientes reales que renovaron
- **4602:** todos los clientes reales que no renovaron.

En el análisis del costo de FN que es calculado por el número de falsos negativos por el costo del valor de renovación de firma anual de $18.40

$$\boldsymbol{FN = FN * 18.40}$$

Y el costo FP que es el cálculo por el número de falsos positivos por el costo de 0.10 dólares por usuario de mensajes por WhatsApp Marketing dando como resultado los datos mostrados en la tabla 4.6.

$$\mathbf{FP = FP * 0.1}$$

TN	FP	FN	TP	Costo FN	Costo FP
2554	4602	0	0	$84676.80	$0

Tabla 4.6:Costos de solución heurística.

A continuación, en la tabla 4.7 y 4.8, detallaremos el análisis de la solución planteada por el modelo predictivo con los distintos umbrales:

Umbral	TN	FP	FN	TP	Costo FN	Costo FP
0.40	1043	6113	0	3379	$0	$611.30
0.41	1043	6113	0	3379	$0	$611.30
0.42	1043	6113	0	3379	$0	$611.30
0.43	1043	6113	0	3379	$0	$611.30
0.44	1048	6108	3	3376	$55.20	$610.80
0.45	1048	6108	3	3376	$55.20	$610.80
0.46	1048	6108	3	3376	$55.20	$610.80
0.47	1190	5966	50	3329	$920	$596.60
0.48	2554	4602	487	2892	$8960.80	$460.20
0.49	2554	4602	487	2892	$8960.80	$460.20
0.50	2566	4590	493	2886	$9071.20	$459

Tabla 4.7: Umbrales y costos de modelo predictivo.

Umbral	Presicion	Recall	F1-score	Accuracy
0.40	0.36	1	0.53	0.42
0.41	0.36	1	0.53	0.42
0.42	0.36	1	0.53	0.42
0.43	0.36	1	0.53	0.42
0.44	0.36	1	0.52	0.42
0.45	0.36	1	0.52	0.42
0.46	0.36	1	0.52	0.42
0.47	0.36	0.99	0.53	0.43
0.48	0.39	0.86	0.53	0.52
0.49	0.39	0.86	0.53	0.52
0.50	0.39	0.85	0.53	0.52

Tabla 4.8: Rendimiento de modelo predictivo y sus umbrales.

Si asumimos que la heurística predice que todos los clientes se quedan (clase 0), entonces:

- **Clase 0 (No abandonan):** Todos los clientes que realmente no abandonan serán verdaderos negativos (TN). Todos los clientes que realmente abandonan serán falsos negativos (FN).

- **Clase 1 (Abandonan):** No habrá verdaderos positivos (TP=0) ni falsos positivos (FP=0), porque nunca se predice que alguien abandona.

Nueva Matriz de Confusión para la Heurística

Con la matriz original:

$$\begin{Bmatrix} \mathbf{2554} & \mathbf{4602} \\ \mathbf{487} & \mathbf{2892} \end{Bmatrix}$$

Podemos calcular los valores:

TN (Clase 0 correctamente clasificada): Todos los clientes de la clase 0.

$$TN = 2554 + 4602 = 7156$$

FN (Clase 1 incorrectamente clasificada como clase 0): Todos los clientes de la clase 1.

$$FN = 487 + 2892 = 3379$$

TP (Clase 1 correctamente clasificada): Ninguno, porque no se predice nunca "abandonan".

$$\boldsymbol{TP = 0}$$

FP (Clase 0 incorrectamente clasificada como clase 1): Ninguno, porque no se predice nunca "abandonan".

$$\boldsymbol{FP = 0}$$

La matriz de confusión sería:

$$\boldsymbol{Heurística\ (Predice\ todos\ se\ quedan)} = \begin{Bmatrix} \mathbf{7156} & \mathbf{0} \\ \mathbf{3379} & \mathbf{0} \end{Bmatrix}$$

Costos Operativos con la Nueva Heurística

Dado que el costo de un falso negativo es 18.40 USD y el de un falso positivo es 0.10 USD, calculamos los costos:

Caso A: No contactar a ningún cliente

Falsos Negativos (FN):

$$FN = 3379$$

$$Costo\ de\ FN = 3379 \cdot 18.40 = \$62177.60$$

Falsos Positivos (FP):

$$FP = 0$$

$$Costo\ de\ FP = 0 \cdot 0.10 = \$0$$

Costo Total

$$Costo\ Total = 62177.60 + 0 = \$62177.60$$

Caso B: Contactar a todos los clientes

Falsos Negativos (FN):

$$FN = 0$$

$$Costo\ de\ FN = 0 \cdot 18.40 = \$0$$

Falsos Positivos (FP):

$$FP = 7500 - 2886 + 493 = 4121$$

$$Costo\ de\ FP = 4121 \cdot 0.10 = \$412.10$$

Costo Total

$$Costo\ Total = 412.10 + 0 = \$412.10$$

Modelo Actual:

Falsos Negativos (FN):

$$FN = 487$$

$$Costo\ de\ FN = 487 \cdot 18.40 = \$8956.80$$

Falsos Positivos (FP):

$$FP = 4602$$

$$Costo\ de\ FP = 4602 \cdot 0.10 = \$460.20$$

Costo Total del Modelo Actual:

$$Costo\ Total = 8956.80 + 460.20 = 9417.00\ USD$$

Pero se realizó un análisis con el top 10% de clientes con probabilidad de desertar, con lo cual se obtuvo la siguiente matriz de confusión:

$$\begin{Bmatrix} \mathbf{0} & \mathbf{525} \\ \mathbf{0} & \mathbf{528} \end{Bmatrix}$$

Falsos Positivos (FP): 525

Falsos Negativos (FN): 0

Modelo (Top 10%)

Falsos Positivos (FP): 525

Falsos Negativos (FN): 0

Costo total

$$(FN \times 18.40) + (FP \times 0.10) = (0 \times 18.40) + (525 \times 0.10) = \$52.50$$

Selección aleatoria (Top 10%)

$$\begin{Bmatrix} \mathbf{0} & \mathbf{728} \\ \mathbf{0} & \mathbf{325} \end{Bmatrix}$$

Falsos Positivos (FP): 728

Falsos Negativos (FN): 0

Costo total:

$$(FN \times 18.40) + (FP \times 0.10) = (0 \times 18.40) + (728 \times 0.10) = 0 + 72.80 = \$72.80$$

Comparación

Estrategia	FN	FP	Costo total
Modelo (Top 10%)	0	525	$52.50
Contactar a todos	0	4121	$412.10
No contactar a nadie	3379	0	$62173.60

Tabla 4.9: Comparativa de estrategias.

A continuación, se explicarán las distintas estrategias comparadas con relación al costo total presentadas por la tabla 4.9.

Modelo (Top 10%): Es la estrategia más eficiente en términos de costo, ya que concentra los esfuerzos en el segmento más probable de abandonar, minimizando tanto Falsos Negativos como el costo asociado a Falsos Positivos.

Contactar a todos: Aunque más caro que el modelo, esta estrategia tiene un costo relativamente bajo debido al pequeño costo de contactar a un cliente ($0.10). Sin embargo, contacta a muchos clientes innecesarios (4,121).

No contactar a nadie: Es con diferencia, la peor estrategia en términos de costo, ya que no mitiga el impacto de los clientes que abandonan ($62,173.60).

El modelo (top 10%) ahorra:

- $359.60 frente a contactar a todos los clientes.
- $62,121.10 frente a no contactar a ningún cliente.

Además, el modelo tiene un costo 87.26% menor que contactar a todos y 99.92% menor que no contactar a nadie, lo que demuestra su capacidad de priorizar correctamente los clientes que más probablemente abandonen y minimizar los costos operativos.

Con este modelo del top 10% se plantea el contacto mediante una campaña de WhatsApp Marketing, focalizada en estos clientes con probabilidad de abandono, se les ofrece un descuento del 15% de descuento sobre el valor unitario de la firma $18.40.

Costos incurridos al aplicar el descuento del 15%:
Costo por contacto

$$\boldsymbol{Costo\ por\ contacto} = Clientes\ contactados \times Costo\ por\ contacto = 1{,}053 \times 0.10 = \$105.30$$

Costo por descuento

$$\boldsymbol{Costo\ del\ descuento} = Clientes\ retenidos \times (VPC \times Descuento) = 528 \times (18.40 \times 0.15) = 528 \times 2.76 = \$1457.28$$

Costo total de campaña

$$\boldsymbol{Costo\ total} = Costo\ por\ contacto + Costo\ del\ descuento = 105.30 + 1{,}457.28 = \$1562.58$$

Pérdida evitada al retener clientes:
Valor de clientes retenidos:

$$\boldsymbol{Valor\ retenido} = Clientes\ retenidos \times VPC = 528 \times 18.40 = \$9715.20$$

Beneficio neto:

$$\boldsymbol{Beneficio\ neto} = Valor\ retenido - Costo\ total = 9{,}715.20 - 1562.58 = \$8152.62$$

Como se evidencia en la tabla 4.10, la aplicación nos brindaría un descuento como inversión estratégica, en donde el descuento del 15% representa una inversión que permite retener a clientes con un valor significativo ($18.40 por cliente). Sin esta intervención, la pérdida total de $9,715.20 USD sería inevitable.

Estrategia	Descuento	Valor de firma	Valor de contacto	Costos	Ingresos	Neto
Modelo top 10%	$2	$18.40	$0.10	$105.30	$9715	$9609.90
Selección aleatoria 10%	$2	$18.40	$0.10	$105.30	$5980	$5874.70
Lifting/Mejora						**$3735.20**

Tabla 4.10: Mejora en la comparación de modelos.

El modelo de aprendizaje automático incrementa notablemente la eficiencia en la selección de clientes, generando un 63% más de ingresos netos en comparación con una selección aleatoria. Esto demuestra que el modelo identifica con éxito a los clientes clave para prevenir la deserción.

Estos hallazgos confirman que el modelo ofrece un retorno positivo de la inversión, haciendo que el desarrollo y mantenimiento de este sea rentable. Además, permite un uso estratégico de los recursos, optimizando el retorno por cada cliente contactado.

A pesar de que el análisis se llevó a cabo con un conjunto reducido de variables y sin incluir datos como edad, sexo, detalles geográficos o profesión (los cuales no están disponibles en el dataset actual), los resultados obtenidos han sido positivos para la empresa. Esto demuestra el potencial de los datos existentes, basados únicamente en facturación, renovaciones y gestiones de soporte al cliente, para generar información valiosa y tomar decisiones estratégicas.

Con base en estos hallazgos, se sugiere a la empresa implementar campañas de marketing por WhatsApp dirigidas específicamente a los clientes con mayor probabilidad de riesgo de abandono, mediante estrategias de retención diseñadas a partir de los patrones detectados en este análisis.

Como pasos futuros, se plantea seguir refinando el modelo para mejorar la precisión en la identificación de clientes, así como implementar análisis de sensibilidad que ajusten el descuento o el valor del contacto, maximizando aún más los resultados obtenidos.
Considerando la incorporación de información adicional al dataset, como datos demográficos, históricos de interacción y comportamientos de uso, para enriquecerlo y fortalecer futuros análisis. Con un conjunto de datos más completo, se podrán desarrollar modelos analíticos más precisos, diseñar campañas personalizadas y maximizar el valor generado por estas iniciativas.

4.7 Consideraciones finales

En este proyecto de predicción de abandono de clientes, varios desafíos clave afectaron el desarrollo y la precisión del modelo. A continuación, se detallan las principales consideraciones que se deben tener en cuenta para interpretar los resultados y planificar futuras mejoras en el desarrollo de este modelo:

Calidad de los Datos: La calidad de los datos fue uno de los mayores obstáculos enfrentados durante el proceso. Se detectaron múltiples problemas relacionados con la consistencia, la cantidad de datos faltantes (valores nulos y en blanco) y la presencia de valores incoherentes (montos de factura aberrantes). Esto obligó a realizar una limpieza intensiva de los datos, lo que podría haber impactado negativamente en la capacidad del modelo para capturar patrones precisos de deserción. Para mejorar futuros proyectos, se recomienda implementar controles de calidad de datos desde la recopilación, asegurando que la información sea precisa y completa. Además de mantener un control de los cambios que se generan en las ofertas de precios por temporadas, para focalizar un estudio en los precios con mayor precisión.

Cantidad de Datos Disponibles: La limitada cantidad de datos disponibles restringió el rendimiento del modelo, debido a la exhaustiva limpieza de datos que se realizó. Lo cual es una desventaja porque las predicciones de churn suelen requerir grandes volúmenes de información para capturar patrones y tendencias de comportamiento de los clientes. Aunque los modelos predictivos lograron resultados razonables, con un conjunto de datos mayor y más representativo se podrían obtener mejoras significativas en la precisión de las predicciones. Se sugiere considerar estrategias como el uso de datos sintéticos o ampliar las fuentes de datos en futuras iteraciones con una mayor apertura al acceso a datos reales.

Fuentes de Datos Utilizadas: El proyecto se basó en tres fuentes de datos: soporte técnico, encuestas de satisfacción y ventas del producto. Aunque estas fuentes proporcionaron una visión limitada del cliente, es probable que no representen completamente todos los factores que inciden en la decisión de un cliente de abandonar. En futuras versiones del modelo, sería valioso incorporar fuentes adicionales de datos más ricos, como interacciones en redes sociales, comportamiento en la web, soportes de llamadas telefónicas y soportes de correo electrónico, lo que podría mejorar la capacidad predictiva del modelo.

Mejora Futuras con Datos Más Complejos: Para mejorar el rendimiento del modelo y su capacidad de generalización, es altamente recomendable integrar otros datasets que ofrezcan una visión más completa del comportamiento del cliente. Datos como interacciones de marketing, hábitos de compra, uso del producto o características demográficas, profesión, edad que podrían ofrecer insights más profundos y permitir una segmentación más precisa.

Trabajos futuros: Como consideraciones finales, es fundamental enfocar esfuerzos futuros en la identificación del comportamiento de deserción de clientes que ocurre uno, dos o más meses después de la fecha de caducidad de la firma electrónica. Actualmente, la falta de datos específicos sobre este comportamiento limita el análisis profundo de cuán importante es el comportamiento del cliente. Al recopilar y analizar esta información, podremos desarrollar estrategias más efectivas para abordar la retención de clientes, anticipar sus necesidades y mejorar los servicios ofrecidos.

4.8 Conclusiones y Recomendaciones

4.8.1 Conclusiones

- El proyecto de predicción de abandono de clientes mediante machine learning ha mostrado un impacto positivo en la retención de clientes y en la optimización de recursos del equipo de renovaciones para enfatizar un seguimiento de los clientes fieles a la marca y enfocarnos en otros métodos de contacto con aquellos que deciden desertar.

- El análisis del comportamiento de los clientes en los meses de noviembre y diciembre reveló que solo se contacta a una pequeña porción de clientes para conocer el motivo de deserción del servicio, de los cuales no se puede generar un primer contacto a través de las campañas de marketing actuales, lo que limita la efectividad de las estrategias de retención.

- Se identificó que, a partir del tercer mes de implementación del modelo, se obtiene un retorno de la inversión (ROI), gracias a la combinación de la reducción de costos en campañas de marketing y el uso eficiente de los recursos humanos asignados al proyecto.

- El uso del modelo predictivo automatizado permite mejorar la precisión en la identificación de clientes con mayor riesgo de abandono, lo que a su vez permite focalizar las intervenciones y reducir costos operativos relacionados con las campañas masivas.

4.8.2 Recomendaciones

- ***Implementar el modelo de churn de manera escalonada:*** Se recomienda seguir utilizando el modelo predictivo en campañas de marketing a lo largo de varios meses, con un análisis constante del retorno de la inversión (ROI) a partir del tercer mes, para asegurar la rentabilidad del proyecto.

- ***Optimizar los esfuerzos de contacto:*** Utilizar las predicciones del modelo para priorizar a los clientes que muestran mayor riesgo de churn, asignando recursos humanos de manera más eficiente para maximizar el impacto en la retención de estos clientes para generar un programa de fidelidad con la marca.

- ***Monitorizar el impacto a largo plazo:*** Continuar midiendo el impacto del modelo no solo en términos de retención de clientes, sino también en costos operativos y financieros, garantizando que el retorno de la inversión sea consistente y mejorando las estrategias según los resultados.

- ***Integrar más fuentes de datos para enriquecer el modelo:*** Dado que actualmente se utiliza información de soporte, encuestas y ventas del producto, se recomienda considerar la incorporación de nuevas fuentes de datos (historial de compras, interacciones en redes sociales, datos demográficos, edad, profesión, entre otros) para mejorar aún más la precisión del modelo.

- ***Capacitar al equipo de renovaciones:*** Con el objetivo de aprovechar al máximo el modelo predictivo, es clave que el equipo de renovaciones reciba capacitación sobre cómo utilizar los insights proporcionados por el modelo para personalizar las estrategias de retención y mejorar la relación con el cliente.

5 Referencias

1. Anitha M A, a. S. (2022). An Efficient Hybrid Classifier Model for Customer. *INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS*, 11-18.

2. Barros, T. &. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. *Education Sciences*.

3. Datta, P. M. (2000). Automated Cellular Modeling and Prediction on a Large Scale. *Artificial Intelligence Review 14*, 485–502 .

4. Eria, K. &. (2018). Systematic Review of Customer Churn Prediction in the Telecom Sector. 7-14.

5. Glory Sam, P. A. (2024). Customer Churn Prediction using Machine Learning Models. *Journal of Engineering Research and Reports*, 181-193.

6. Gore, S. a. (2023). Customer Churn Prediction using Neural Networks and SMOTE-ENN for Data Sampling. *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*, (págs. 1-5). India.

7. Imani M, A. H. (2023). Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis. *Technologies*.

8. IRFAN ULLAH, B. R. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*.

9. Jaishankar Ganesh, M. J. (2000). Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers. *Journal of Marketing*, 65-67.

10. Kedar Potdar, T. S. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 7-9.

11. Lalwani, P. M. (2022). Customer churn prediction system: a machine learning approach. *Computing 104*, 271–294.

12. Levent ÇALLI, S. K. (2022). Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry. *Academic Platform Journal of Engineering and Smart Systems (APJESS)* , 115-123.

13. Louis Geiler, S. A. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, 217-242.

14. Louis Geiler, S. A. (2022). An effective strategy for churn prediction and customer profiling. *Data & Knowledge Engineering*.

15. Marcel Karnstedt, M. R. (2011). The Effect of User Features on Churn in Social Networks. *Proceedings of the 3rd International Web Science Conference*.

16. Mittal, B. &. (1998). Why do customers switch? The dynamics of satisfaction versus loyalty. *Journal of Services Marketing*, 177–194.

17. Mohammed, R. R. (2020). Mohammed, R., Rawashdeh, JMachine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS).*

18. Mozer, M. a. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 690-696.

19. Nicolas Glady, B. B. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 402-411.

20. PIEW DATTA, B. M. (2000). Automated Cellular Modeling and Prediction on a Large Scale. *Artificial Intelligence Review* , 485-502.

21. Ping Jiang, Z. L. (2024). Profit-driven weighted classifier with interpretable ability for customer churn prediction. *Omega*.

22. Praveen Lalwani, M. K. (2021). Customer churn prediction system: a machine learning approach. *Computing*.

23. Pristyanto, A. Y. (2022). Machine Learning Models for Classifying Imbalanced Class Datasets Using Ensemble Learning. *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 648-653.

24. Richard J. Oentaryo, E.-P. L. (2012). Collective Churn Prediction in Social Network. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

25. S. Agrawal, A. D. (2018). Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning. *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 1-6.

26. Sam, G. &. (2024). Customer Churn Prediction using Machine Learning Models. *Journal of Engineering Research and Reports.*, 181-193.

27. Seyed Hossein Iranmanesh, M. H. (2019). Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application . *Proceedings of the International Conference on Industrial Engineering and Operations Management*.

28. Ullah, I. &. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access.*

29. Wouter Verbeke, D. M. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 2354-2364.

30. Xie, Y. L. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 5445–5449.

31. Yap, B. W. (2013). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Proceedings of the First International Conference on Advance*.

32. Zeithaml, V. &. (1996). The Behavioral Consequences of Service Quality. *Journal of Marketing.*