UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE CLASIFICACIÓN DE MOTIVOS DE FUGA POR PROBLEMAS DE RED EN UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PATRICIO ARMANDO PÉREZ DELGADO

PROFESOR GUÍA:
NICOLÁS CISTERNAS GONZÁLEZ

MIEMBROS DE LA COMISIÓN:
MARIA FERNANDA VARGAS COURBIS
BLAS DUARTE ALLEUY

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: PATRICIO ARMANDO PÉREZ DELGADO
FECHA: 2024
PROF. GUÍA: NICOLÁS CISTERNAS GONZÁLEZ

MODELO DE CLASIFICACIÓN DE MOTIVOS DE FUGA POR PROBLEMAS DE RED EN UNA EMPRESA DE TELECOMUNICACIONES

El presente trabajo de título tiene como objetivo desarrollar un modelo de clasificación para predecir si un cliente de una compañía de telecomunicaciones tiene alto riesgo de abandonar el servicio debido a problemas en la red que generan una mala experiencia de usuario. La empresa que se analiza es Entel, el mayor operador móvil de Chile. En donde la portabilidad es uno de los principales canales de fuga de clientes. Dado que la ley de portabilidad ha eliminado las barreras de salida, no se registra el motivo por el cuál un cliente decide portarse a otra compañía, solo se da aviso posteriormente, por lo que no hay una estrategia enfocada a la causa de fuga. Para ello se busca detectar los casos de fugas que se deben específicamente a fallas técnicas mediante el análisis de métricas de experiencia de red y su correlación con motivos de salida declarados en encuestas posteriores.

El proyecto se enmarca en el campo de la ciencia de datos, utilizando técnicas de aprendizaje supervisado y en específico modelos de clasificación binaria. La metodología seguida fue CRISP-DM, que estructura el desarrollo en fases de comprensión del negocio, entendimiento de datos, preparación de datos, modelamiento, evaluación e implementación. Se utilizó información de clientes que se portaron de Entel entre abril y octubre del 2023, con variables relativas a su experiencia previa en la red como intensidad de señal, cambios de tecnología, minutos cursados y tráfico de internet. Estos datos se vincularon con encuestas posteriores que capturan la motivación detrás de su partida.

Se probó una variedad de algoritmos de clasificación binaria. El modelo CatBoost demostró la mejor capacidad para distinguir clientes con alta y baja propensión de abandono técnico según la métrica AUC, que alcanzó un máximo de 0.71 en set de prueba. Se analizó a profundidad el performance en términos de métricas como precisión, exactitud y curva ROC y Lift. También se analizó la interpretabilidad de variables con la técnica SHAP Values y se comparó el funcionamiento de las predicciones de modelo versus la distribución de datos reales.

Los resultados permiten priorizar de forma proactiva mejoras en la calidad e infraestructura de la red móvil de Entel en las zonas geográficas donde se concentran los usuarios con mayor probabilidad pronosticada de experimentar problemas técnicos severos que deriven en su portabilidad, ayudando a enfocar las acciones de retención.

Tabla de Contenido

Índice de tablas

Índice de ilustraciones

Capítulo 1: Antecedentes Generales

1.1. Contextualización Industria

La industria de las telecomunicaciones en Chile se destaca por ser compacta, solo las cuatro principales compañías dominan más del 90% del mercado actual (Subtel, 2023). Existe una amplia gama y variedad de planes, tarifas y paquetes de servicios que van desde telefonía móvil y fija hasta servicios de internet y televisión que están diseñados para atraer diferentes segmentos de clientes. Aunque está dominada por 4 grandes empresas, estas compiten agresivamente entre sí para ganar clientes y cuota de mercado. Cada empresa intenta ofrecer mejores tarifas, planes y servicios que las demás para captar suscriptores.

La industria de telecomunicaciones en Chile ha experimentado una gran transformación impulsada por la innovación tecnológica, el aumento de la cobertura de redes y los esfuerzos en inclusión digital. La constante expansión de redes fijas y móviles hacia zonas menos densamente pobladas ha abierto el acceso a servicios de telecomunicaciones a más segmentos de la población. Esto amplía el mercado disponible aumentando las posibilidades de elección de las personas. Esto ha derivado en una mayor competencia que se ha traducido en beneficios concretos para los consumidores. La introducción de nuevas tecnologías como 4G y fibra óptica han permitido el desarrollo de planes más flexibles y enfocados en datos móviles en lugar de minutos, adaptándose mejor a los patrones de consumo actuales. En 2020 una medición global posicionó a Chile en el lugar 20° entre los países con el precio promedio más bajo de 1 GB (América Economía, 2020).

A marzo 2023 el sector alcanza los 58,9 millones de servicios de telecomunicaciones desplegados en la modalidad de suscripción, evidenciando el crecimiento y alcance de la industria en el país (Subtel, 2023).

La telefonía móvil, servicio en el que un cliente accede a llamadas y conexión a internet a través de su teléfono, se posiciona como el servicio de telecomunicaciones con mayor penetración en la población, en marzo del 2023 abarca más del 50% de los servicios activos en Chile. La telefonía móvil alcanzó una significativa penetración de 133,3 abonados por cada 100 habitantes. Esta métrica refleja la extensión generalizada de los servicios de telefonía móvil en la población, indicando un acceso extendido más allá de una línea telefónica por habitante. Los servicios de pospago alcanzan 70% del total de abonados, superando considerablemente a los abonados de prepago. Esta inclinación sugiere una transición significativa hacia compromisos a largo plazo, con implicaciones relevantes para la estabilidad y fidelización de la base de usuarios (Subtel, 2023).

Normalmente los servicios de telefonía móvil, la telefonía fija, internet hogar y televisión son ofrecidos en paquetes y con promociones por una mayor cantidad de clientes en la cuenta. Es esencial reconocer que el buen funcionamiento del ecosistema de telecomunicaciones depende de la calidad de todos los servicios. En este contexto, el internet emerge como un pilar fundamental para garantizar el óptimo rendimiento de diversos servicios. La oferta de paquetes que integran servicios como televisión, internet por fibra óptica y planes de internet hogar móvil se presenta como una opción atractiva de entretenimiento. Asimismo, se observa un crecimiento significativo en la demanda de servicios de valor agregado, lo que ha llevado a todas las compañías a ofrecer estos servicios. Por ejemplo, la opción de cargar el costo de los servicios como Netflix y Spotify a la factura telefónica reflejan un cambio hacia opciones digitales más flexibles y convenientes, evidenciando la interconexión y dependencia de todo el ecosistema de telecomunicaciones (Subtel, 2018).

La regulación también ha desempeñado un papel en la evolución de los precios. Las autoridades gubernamentales han tomado medidas para fomentar la competencia y evitar prácticas que podrían conducir a tarifas elevadas, un ejemplo es la ley de la Portabilidad Numérica N°20.471, promulgada en Chile en 2010, que otorga a los usuarios de telefonía la libertad de cambiarse de compañía sin tener que cambiar su número telefónico.

La portabilidad numérica ha representado tanto desafíos como oportunidades para las empresas de telecomunicaciones en Chile. Por un lado, les ha exigido redoblar esfuerzos para mejorar su calidad de servicio y experiencia al cliente. Al tener los usuarios mayor libertad para cambiarse de operadora, las compañías deben competir fuertemente en estos aspectos para retener abonados y no verse afectadas por una ola de portabilidades de salida. Esto ha obligado a las operadoras a, invertir en soporte técnico y servicio al cliente de primer nivel, desarrollar políticas de compensación y satisfacción al cliente para aquellos que experimenten algún tipo de problema en el servicio y perfeccionar la calidad de sus redes y niveles de cobertura para entregar conectividad confiable y evitar quejas sobre interrupciones del servicio que deriven en cambio de operadora. En su informe de 2012, la Subsecretaría de Telecomunicaciones (SUBTEL) de Chile concluyó que la portabilidad numérica había tenido un impacto positivo en el sector. El informe encontró que la ley había contribuido a reducir los precios de los servicios de telefonía móvil en un 20%. (Subtel, 2012).

También ha sido una oportunidad para captar clientes insatisfechos de la competencia, aprovechando falencias puntuales en su servicio para atraer a estos abonados descontentos. Desde el inicio del sistema y hasta el 30 de septiembre de 2023 se han portado más de 34 millones de números (34.520.630), de los cuales el 95,6% (32.998.972) corresponden a portaciones de números móviles. (Subtel, 2023).

Como se indica en la ilustración 1, a junio de 2023, cuatro operadores lideran el mercado de telefonía móvil. Entel, Movistar, WOM y Claro controlan conjuntamente el 98,3% de la participación, medida en cantidad de clientes. Entre ellos, Entel es el líder con una participación de mercado del 32,8%, seguido por Movistar con el 25,7%. WOM se posiciona como el tercer operador más grande con una participación del 21,5%, mientras que Claro ocupa el cuarto lugar con el 18,3% de participación (Subtel, 2023).

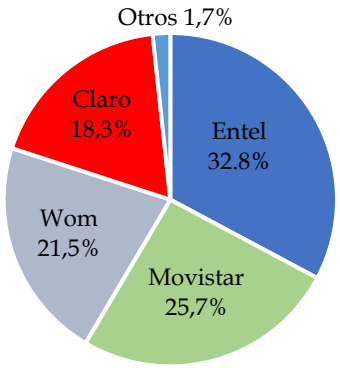

Ilustración 1: Market Share de telefonía móvil en cantidad de abonados.

La columna vertebral de las telecomunicaciones móviles se fundamenta en la operación de los conocidos "sitios de telecomunicaciones", que son los emplazamientos físicos donde está toda la infraestructura técnica de redes, estos puntos desempeñan un papel crucial en el entramado de la comunicación móvil al permitir que se transmita información mediante ondas de radio. La planificación cuidadosa de la ubicación de estos sitios no es un detalle menor, sino un aspecto estratégico. La distribución se lleva a cabo considerando factores clave como la densidad poblacional, la topografía del terreno y los patrones de uso de los usuarios. Esto no solo asegura una cobertura extensiva, sino también una conectividad eficiente y equitativa, adaptada a las necesidades específicas de la población en diferentes áreas geográficas (Entel S.A., 2020).

En el contexto de las telecomunicaciones, la fuga de clientes representa el fenómeno de usuarios que cancelan o dan de baja su suscripción. La clasificación de los clientes que abandonan la compañía incluye tanto el "churn voluntario", cuando el cliente decide dar de baja el servicio o cambiar a otra compañía, como el "churn involuntario", donde se da de baja el servicio debido a falta de pago o fraude. La clasificación principal de fuga de clientes es el "Port Out", que se refiere al proceso mediante el cual un cliente decide cambiar de un proveedor de servicios a otro, llevándose consigo su número de teléfono existente. Este fenómeno es una forma de churn voluntario, ya que el cliente toma la decisión activa de cambiar de proveedor.

La falta de una gestión efectiva de las fugas de clientes puede acarrear diversos problemas para las empresas. Esto incluye consecuencias perjudiciales para los ingresos y la imagen de marca,

señalando posibles insatisfacciones con los servicios, tarifas poco competitivas o una experiencia deficiente para el cliente. Para contrarrestar las pérdidas de clientes, se requiere la adquisición de nuevos usuarios, lo que implica costos de marketing y promoción adicionales, aumentando los gastos de las empresas. En Chile las compañías abordan esta problemática desde distintas perspectivas, sus acciones han ido desde ofrecer incentivos para retener clientes, como descuentos, promociones o programas de lealtad, hasta la mejora de su infraestructura de red para satisfacer las necesidades de sus clientes.

1.2. Contextualización Empresa

Entel, fundada en Chile en 1964 como una empresa estatal de telecomunicaciones, tiene la misión de proporcionar servicios a nivel nacional. A lo largo de su historia, Entel ha demostrado un compromiso constante con la inversión en infraestructura y tecnología, consolidando así su reputación de calidad en el mercado chileno de las telecomunicaciones. En la actualidad, dos de sus pilares estratégicos son la experiencia distintiva y la infraestructura moderna y robusta. La experiencia distintiva significa ofrecer a sus clientes una experiencia de servicio que sea superior a la de sus competidores. Esto incluye factores como la calidad de la conectividad, la atención al cliente, la innovación y la variedad de servicios ofrecidos. La infraestructura moderna y robusta significa contar con una red de telecomunicaciones que sea capaz de satisfacer las necesidades actuales y futuras de los clientes. Esto incluye factores como la cobertura, la capacidad y la tecnología utilizada. En este sentido empresa ha expandido y mejorado continuamente su red, implementando tecnologías avanzadas como 3G, 4G y actualmente 5G. Este enfoque ha posicionado a Entel como el operador de telefonía móvil más grande de Chile, con la mayor cantidad de clientes y la mayor participación de mercado.

Al cierre del primer semestre del 2023, Entel posee cerca de 8.7 millones de clientes de telefonía móvil, incluyendo prepago y pospago, lo que representa el 32,8% del mercado (Entel S.A., 2023). La presencia de Entel en Chile es muy amplia. La compañía tiene cobertura a lo largo todo el territorio nacional, incluyendo zonas rurales y aisladas. Esto le permite ofrecer sus servicios a una amplia gama de clientes, desde personas que viven en ciudades hasta personas que viven en zonas rurales.

Una red confiable y de alto rendimiento es esencial para conservar a los clientes existentes y atraer nuevos usuarios. Esto se basa en diversas razones clave. En primer lugar, una red móvil que ofrece llamadas claras, conexiones estables, velocidades de datos rápidas y una amplia cobertura geográfica contribuye a una experiencia de usuario satisfactoria. Los usuarios contentos son más propensos a permanecer y continuar utilizando los servicios. Según análisis de encuestas realizadas por Entel a clientes que dieron de baja su suscripción recientemente, aquellos que se declaran satisfechos dejan la compañía en su mayoría por razones externas a la calidad del servicio, mientras que

aquellos que se declaran insatisfechos, casi un 50% de ellos declara haberlo hecho por la mala calidad del servicio.

En ese sentido en un entorno altamente competitivo la calidad de la red móvil se convierte en un diferenciador esencial. Las compañías con redes móviles que no presentan problemas, desconexiones y que son rápidas tienen una ventaja significativa al retener a sus clientes y atraer a nuevos, ya que los usuarios están dispuestos a pagar por una experiencia mejorada. La fiabilidad y eficiencia de una red móvil superior se traducen directamente en un servicio más consistente y de calidad para los usuarios, fomentando así la lealtad. Asimismo, la confianza en una red confiable alimenta la fidelidad de los clientes. Aquellos que confían en conexiones estables tienen más probabilidades de permanecer en la misma red. Una red deficiente puede resultar en llamadas interrumpidas o conexiones lentas, lo que genera frustración y podría llevar a los usuarios a buscar alternativas, es decir utilizar el servicio de otra compañía.

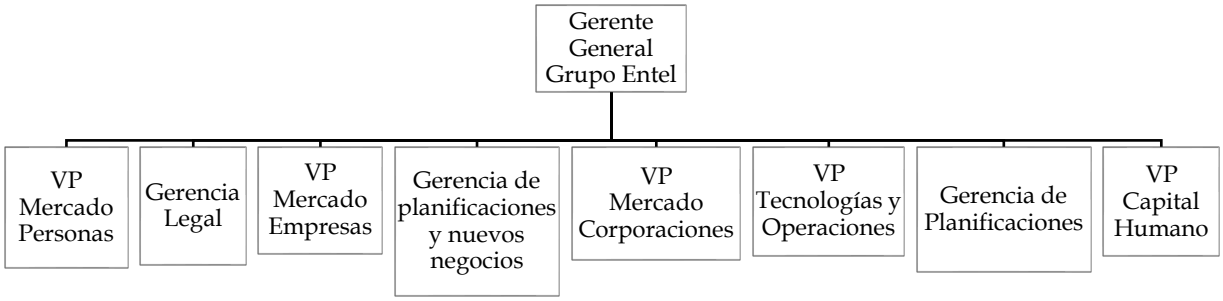

Ilustración 2: Organigrama Gerencias y Vicepresidencias Entel Chile S.A.

Entel ha establecido otros dos pilares estratégicos fundamentales: Innovación y adaptación y organización y cultura ágil. Estos principios han sido cruciales en la estructuración de su actual administración, como se detalla en de la ilustración 2, con segmentos de mercado específicos: Mercado Personas, Mercado Empresas y Mercado Corporaciones. Esta organización integral, bajo el liderazgo de la Gerencia General, está diseñada para abordar eficientemente las diversas necesidades de los clientes, fomentando una interacción cercana. La adopción de una cultura ágil es clave, brindando mayor flexibilidad y capacidad de respuesta a las demandas cambiantes del mercado. La VP de Mercado Personas administra la mayor cantidad de clientes, al incluir todos los servicios contratados por personas naturales en donde se incluye el de telefonía móvil.

La compañía trabaja en convertirse a una empresa “data driven” (Entel S.A., 2023), que significa utilizar datos y análisis para guiar decisiones y estrategias comerciales, en lugar de depender exclusivamente de intuiciones o experiencias pasadas. La clave para alcanzar este objetivo radica en la gestión de datos y la analítica, por eso la empresa ha establecido equipos de inteligencia de negocios que se centran en segmentos de mercado específicos y se adaptan según el tipo de análisis

necesario. Este enfoque garantiza que la toma de decisiones y las estrategias comerciales estén respaldadas por información precisa y relevante. Los equipos de Analytics, se enfocan en proporcionar información valiosa que impacte directamente en la mejora de la gestión y la continuidad operativa. Estos análisis también benefician a los clientes internos, contribuyendo así a una toma de decisiones más informada, precisa y ágil. Un ejemplo concreto es el equipo de Experiencia de Clientes del Mercado Personas, que se dedica a analizar y comprender la relación entre la experiencia de usuarios de servicios de internet y telefonía móvil y su comportamiento, abordando aspectos como la contratación, cancelación de servicios y patrones de uso.

Capítulo 2: Justificación del Proyecto

Este proyecto cumple un rol clave al abordar un desafío complejo en la industria de telecomunicaciones que es dilucidar la relación entre la experiencia del usuario con la red móvil y la probabilidad de fugas por insatisfacción técnica. Como se ha nombrado anteriormente, en un entorno donde los consumidores cuentan con diversas opciones y alternativas en proveedores de servicios, la innovación tecnológica y la calidad del servicio son factores primordiales. Según datos internos la compañía pierde entre 50 y 55 mil clientes cada mes a través de la portabilidad.

Entel realiza cada mes encuestas telefónicas a un grupo de clientes que se han portado a otra compañía recientemente. Se les consulta por el motivo de esta decisión y el nivel de satisfacción en una escala 1-7 del servicio que estaban recibiendo. Si bien la calidad de la red incide en la satisfacción del cliente, no siempre es el único ni principal determinante de la decisión de portarse a otra compañía. Existen múltiples motivaciones detrás de esta acción, desde ofertas comerciales más convenientes hasta malas experiencias en atención al usuario como puede ser una facturación errónea.

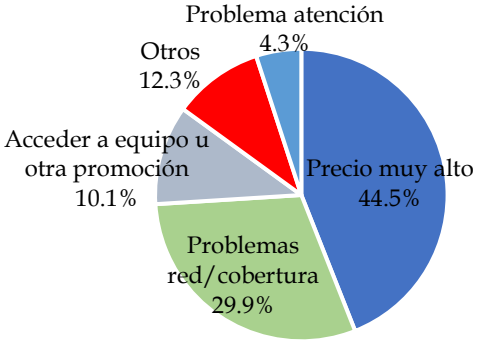

Ilustración 3: Distribución de motivos de fuga declarados en encuesta de portabilidad.

En la ilustración 3 detalla la distribución de motivos de fuga para clientes que se portaron a otra compañía desde Entel. Tiene sentido un alto porcentaje del factor precio, ya que como se declara

anteriormente la competencia en Chile a nivel de precios es alta. Si bien es posible anticipar un alza de fugas por motivos de precio ante ajustes de tarifas como por ejemplo por IPC, no es tan fácil identificar donde o cuando habrá fugas por motivos de red o cobertura.

La ley de portabilidad permite que un cliente se de baja sin notificar a la compañía de origen, lo que limita la capacidad de la empresa para tomar medidas preventivas y efectivas en contra de este fenómeno. Esto limita focalizar iniciativas de optimización sobre áreas geográficas o segmentos de mercado más afectados por una calidad percibida deficiente al no contar con métricas ni patrones claros para identificarlas.

Contar con información sobre cuántos y qué clientes efectivamente mencionarían deficiencias técnicas entre sus razones para abandonar la empresa, puede entregar información valiosa. Saber si existen patrones frecuentes vinculados a interrupciones de servicio, lentitud de navegación o zonas con baja o nula señal, permite detectar oportunidades de mejora en la infraestructura de red.

Abordar estas debilidades contribuiría a aminorar la insatisfacción de un segmento de usuarios y, en conjunto con otras acciones enfocadas en el servicio, podría reducir las tasas actuales de portabilidad por motivos de red o cobertura. Esto incluye problemas como baja cobertura, velocidades de conexión lentas o interrupciones frecuentes que podrían afectar la calidad del servicio. Si los problemas persisten y no se abordan adecuadamente, los usuarios seguirán experimentando inconvenientes y serán más propensos a ver otras opciones para contratar el mismo servicio.

La información de las encuestas de portabilidad puede ser utilizada como etiquetas para el entrenamiento de un modelo de machine learning que discrimine los motivos de fuga y los resultados de este modelo se pueden extrapolar a los clientes actuales, estas predicciones se entienden como la propensión de un cliente de portarse a otra compañía por un motivo de problemas de red.

El propósito principal que se persigue con la implementación de este modelo es reducir la fuga por causas técnicas, anticipándose a través de alertas que permiten una gestión preventiva sobre los usuarios con mayor probabilidad de abandonar a Entel expresamente por fallas atribuibles a cortes de servicio, congestión de red, cobertura deficiente u otros motivos de índole técnico. La detección temprana permite desplegar iniciativas tanto a nivel de red como de satisfacción al cliente, para recuperar la preferencia ante de que se ejecute la portabilidad del número hacia otra operadora.

Capítulo 3: Rol del Estudiante

El estudiante, desde su perfil de Ingeniero Civil Industrial con conocimientos en estadística, análisis de datos y compresión de procesos de negocio, desempeñará el rol de Científico de Datos (Data Scientist) en el equipo de Analytics de Experiencia de Clientes en la VP de Mercado Personas de Entel. Su labor principal será desarrollar un modelo estadístico capaz de identificar si las fugas de clientes están relacionadas con problemas de red y cobertura.

Sus principales responsabilidades serán:

- Liderar el diseño e implementación de un modelo analítico para detectar patrones entre portabilidades y falencias técnicas reportadas por los propios clientes. Procesar y explorar conjuntos de datos de experiencia de clientes, reclamos técnicos, métricas de red, geolocalización de eventos en la red y portabilidades.

- Trabajar coordinadamente con equipos técnicos y comerciales para interpretar resultados del modelo, y para enfocar acciones preventivas en las áreas con mayor propensión de fuga técnica, con la finalidad de identificar correlaciones y derivar insights sobre los principales motivadores de abandono asociados a mal servicio técnico en áreas geográficas específicas.

El alcance del estudiante se centra en la detección de drivers analíticos, modelo predictivo y recomendaciones sobre esta problemática, apoyando con su capacidad de análisis de los datos más que en la implementación de soluciones técnicas, por ello debe traducir estos hallazgos en una herramienta predictiva que permita anticipar nuevas fugas atribuibles a problemas técnicos en zonas críticas.

Capítulo 4: Objetivos

4.1. Objetivo General

Diseñar e implementar un modelo de clasificación binaria para predecir si un cliente presenta alto riesgo de portarse a otra compañía telefónica debido a una mala experiencia de servicio en la red móvil, con la finalidad de identificar a los clientes con alta propensión de fuga por motivos técnicos atribuibles a la red y las variables asociadas al problema.

4.2. Objetivos Específicos

Con el propósito de alcanzar el objetivo general, se formulan los siguientes objetivos específicos.

1. Procesar y analizar los datos históricos de portabilidad y registros de red para identificar patrones predictivos de insatisfacción de cliente y calidad del servicio técnico.
2. Diseñar un conjunto de variables explicativas a partir de los datos disponibles, que capturen señales de insatisfacción técnica e incorporarlas al modelo.
3. Entrenar y evaluar algoritmos de clasificación binaria para estimar la probabilidad de ocurrencia de fuga por mal servicio técnico para cada cliente e identificar las variables más relevantes para modelar el problema.
4. Seleccionar el modelo con mejor desempeño en métricas de discriminación para implementar el modelo resultante para su uso por canales técnicos y comerciales en la gestión preventiva de fugas atribuibles a fallas de red.

Capítulo 5: Alcances

El objetivo es desarrollar un modelo capaz de predecir si el motivo por el cual un cliente se cambió a otra compañía está relacionado con una mala experiencia en la red móvil de Entel. Para lograr esto, se utilizarán exclusivamente herramientas de clasificación binaria, descartando el uso de técnicas de aprendizaje no supervisado. El modelo se utilizará para predecir áreas geográficas específicas donde las fugas debido a problemas de red son más probables, lo que permitirá a Entel proponer mejoras en esas zonas.

La información acerca de los motivos de fuga declarados por clientes será obtenida únicamente mediante encuestas de portabilidad. Estas encuestas se llevan a cabo mensualmente con clientes que han decidido cambiar de proveedor se telefonía móvil en donde se les consulta específicamente sobre los motivos que impulsaron su decisión también su nivel de satisfacción general con el servicio. Sin embargo, es fundamental reconocer la posible presencia de sesgo en estas respuestas. Los clientes pueden proporcionar información sesgada o influenciada por la interpretación personal de su experiencia.

La generación un set de datos para el entrenamiento del modelo involucrará fuentes de datos internas. Estas fuentes abarcarán registros de llamadas, conexiones a antenas y tráfico en las redes 3G y 4G. Se integrarán de manera estructurada para proporcionar una visión integral de la experiencia del usuario. El cruce de datos entre la experiencia de los clientes y las encuestas de portabilidad genera un conjunto de datos que abarca el período comprendido desde abril de 2023 hasta octubre de 2023.

Este proyecto se enmarca en el servicio de telefonía móvil de Entel, por lo que los resultados no son extrapolables a otros servicios ofrecidos por la compañía. Aunque se está desarrollando un modelo predictivo, el uso comercial de este modelo no recae en la responsabilidad del estudiante. Además, queda fuera de los alcances el estudio de costos de infraestructura para implementación de la solución y la responsabilidad posterior por gestión comercial o técnica derivada de las predicciones.

Capítulo 6: Marco Conceptual

Este proyecto está arraigado en el campo de la Ciencia de Datos una rama de la Ingeniería que se enfoca en el análisis de datos para la toma de decisiones.

6.1. Ciencia de Datos y Machine Learning

La Ciencia de Datos es un campo interdisciplinario que se centra en el estudio y la aplicación de métodos, procesos y sistemas para extraer conocimiento y perspectivas útiles a partir de datos en diversas formas. Combina elementos de estadísticas, matemáticas, informática y dominio específico para analizar y comprender fenómenos complejos. Los científicos de datos utilizan una variedad de técnicas, como el análisis exploratorio de datos y el machine learning para examinar conjuntos de datos grandes y complejos. El proceso típico de ciencia de datos implica la recopilación de datos, su limpieza y preparación, la aplicación de modelos analíticos, y la interpretación y comunicación de los resultados (García, 2018).

El objetivo final es descubrir patrones, tendencias y relaciones que puedan ser útiles para la toma de decisiones en diversos campos, desde negocios y finanzas hasta ciencia y salud. La Ciencia de Datos se ha vuelto cada vez más importante en la era de la información, ya que las organizaciones buscan aprovechar al máximo la gran cantidad de datos disponibles para mejorar sus operaciones y tomar decisiones más informadas. La disciplina continúa evolucionando con el desarrollo de nuevas tecnologías y enfoques analíticos.

El machine learning desempeña un papel crucial en la ciencia de datos al permitir que los sistemas aprendan automáticamente patrones y relaciones en los datos, haciendo posible la toma de decisiones y la generación de predicciones sin una programación explícita. Muchos algoritmos y métodos de esta área se basan en conceptos matemáticos y estadísticos para realizar tareas de predicción y

aprendizaje a partir de datos utilizando conceptos como álgebra lineal, cálculo, estadísticas y probabilidad. Se divide en distintos enfoques según la naturaleza de los datos los objetivos deseados. (Alloghani, 2020)

Aprendizaje No Supervisado: En este caso, el modelo se entrena con datos no etiquetados, y su objetivo es descubrir patrones y estructuras dentro de los datos, como la agrupación o reducción de dimensionalidad.

Aprendizaje Supervisado: El modelo se entrena utilizando un conjunto de datos etiquetado, donde las entradas y las salidas deseadas ya están especificadas. El objetivo es aprender la relación entre las entradas y las salidas para poder hacer predicciones precisas en nuevos datos. Los modelos supervisados se dividen comúnmente en dos categorías principales: modelos de regresión y modelos de clasificación. Los modelos de regresión se utilizan cuando la variable de interés es continua, mientras que los modelos de clasificación se aplican cuando se busca predecir la pertenencia a una categoría específica.

6.2. Modelos de clasificación binarios

Los modelos de clasificación binaria son una categoría específica de modelos supervisados diseñados para resolver problemas en los que la variable de interés tiene dos posibles resultados o clases mutuamente excluyentes. Esta formulación se adapta a situaciones donde la tarea es predecir si una observación pertenece a una categoría particular o no. Algunos de estos modelos pueden extenderse a clasificación multiclase o regresión (Mahesh, 2020).

Se listan a continuación algunos algoritmos de clasificadores binarios y una breve descripción de su funcionamiento:

1. **K - Neighbors (K - Vecinos más Cercanos):** Clasifica una observación según la mayoría de la clasificación de sus k vecinos más cercanos en el espacio de características. En machine learning, el "espacio de características" es un concepto que se refiere al conjunto de todas las características o variables que se utilizan para describir un objeto o una observación en un conjunto de datos. Cada observación se representa como un punto en este espacio, y cada dimensión corresponde a una característica específica.
2. **Regresión Logística:** Un clasificador lineal que utiliza la función logística para modelar la probabilidad de pertenecer a una clase en problemas de clasificación binaria o multiclase. La función logística, también conocida como función sigmoide, es una función matemática que mapea cualquier valor real a un rango entre 0 y 1.

3. **Linear Discriminant Analysis (Análisis Discriminante Lineal):** Busca el hiperplano que mejor separa las clases maximizando la distancia entre las medias y minimizando la varianza dentro de cada clase. En un problema de clasificación binaria, el hiperplano es un subconjunto del espacio de características que actúa como una superficie de decisión. Este hiperplano es (n-1)-dimensional, donde "n" es la cantidad de dimensiones del espacio de características. En otras palabras, si estás trabajando en un espacio de características tridimensional, el hiperplano sería bidimensional y se representaría como un plano en el espacio

4. **Decision Tree (Árbol de Decisión):** Construye un árbol donde cada nodo representa una decisión basada en las características, dividiendo de manera jerárquica el espacio de características. Es versátil y fácil de entender, siendo útil para problemas complejos.

5. **Random Forest (Bosque Aleatorio):** Este conjunto de árboles de decisión mejora la precisión y generalización al combinar múltiples modelos. Cada árbol con su resultado vota por la predicción y el resultado final se determina por votación mayoritaria, a esta técnica se le llama Bagging.

6. **Extremely Randomized Trees:** Similar a Random Forest, utiliza múltiples árboles de decisión, pero con la particularidad de seleccionar nodos de división de forma extremadamente aleatoria. Esto incrementa la diversidad entre los árboles, mejorando la robustez del modelo.

7. **Gradient Boosting:** Un algoritmo que construye un modelo fuerte de manera secuencial, corrigiendo los errores del modelo anterior, esta técnica se llama Boosting. Da más peso a las observaciones mal clasificadas, mejorando la precisión y permitiendo la captura de patrones más complejos en los datos. XGBoost, LightGBM y CatBoost son implementaciones avanzadas de Gradient Boosting que han sido diseñadas para abordar diferentes desafíos y mejorar la eficiencia y la precisión en diversas situaciones (Prokhorenkova, 2018).

En el contexto de la clasificación binaria, es crucial comprender dos conceptos fundamentales que desempeñan un papel central en la toma de decisiones del modelo:

1. **Probabilidad de Predicción:** La probabilidad de predicción es una estimación continúa proporcionada por el modelo, indica la probabilidad de que una observación pertenezca a la clase positiva. Esta medida varía entre 0 y 1, donde 0 representa una probabilidad baja y 1 una probabilidad alta de pertenencia a la clase positiva.

2. **Threshold (Umbral):** El threshold es el valor límite que determina cómo se clasifica una observación después de que el modelo ha calculado la probabilidad de predicción. Si la probabilidad es igual o mayor que el threshold, la observación se clasifica como positiva; de lo contrario, se clasifica como negativa (Freeman, 2008).

6.3. División de Datos para Entrenamiento

División en Conjuntos de Entrenamiento y Prueba: En este enfoque, el conjunto de datos se divide aleatoriamente en dos partes principales: un conjunto de entrenamiento, que se utiliza para entrenar el modelo, y un conjunto de prueba, que se reserva para evaluar la capacidad de generalización del modelo en datos no vistos. La asignación típica es del 80% de los datos al conjunto de entrenamiento y el 20% al conjunto de prueba.

Validación Cruzada: Se divide el conjunto de datos en k partes o pliegues y realiza k iteraciones de entrenamiento y evaluación. Cada iteración utiliza un pliegue diferente como conjunto de prueba, proporcionando una estimación más robusta del rendimiento del modelo. La división puede ser estratificada para asegura que la proporción de clases en cada pliegue sea similar a la proporción en el conjunto de datos original, siendo especialmente útil en situaciones de desbalance de clases (validación cruzada estratificada). Esto garantiza que cada clase esté representada de manera equitativa en todos los pliegues, contribuyendo a una evaluación más precisa y justa del modelo.

6.4. Métricas de Desempeño

Las métricas de desempeño son medidas cuantitativas que se utilizan para evaluar la calidad y el rendimiento de un modelo predictivo o clasificador. Estas métricas proporcionan información sobre la precisión y la capacidad de generalización y otros aspectos importantes del modelo. Para calcular las métricas de desempeño, se utiliza la comparación entre la etiqueta real y la predicha para el modelo. Generalmente se comparan las métricas de desempeño en el set de entrenamiento con las del set de testeo o validación, según la estrategia de división de datos que se utilice (Vuk, 2006).

Algunas métricas de desempeño comunes son:

1. **Matriz de Confusión:** Una tabla que muestra el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Es útil para visualizar el desempeño del modelo de manera más detallada.

		Valores Reales	
		Positivo (1)	Negativo (0)
Valores Predichos	Positivo (1)	Verdadero Positivo	Falso Positivo

	Negativo (0)	Falso Negativo	Verdadero Negativo

Ilustración 4: Matriz de confusión

2. **Exactitud (Accuracy):** La proporción de predicciones correctas con respecto al total de predicciones. Se calcula como:

$$Accuracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total\ de\ Predicciones}$$

3. **Precisión (Precision):** La proporción de verdaderos positivos (instancias positivas correctamente clasificadas) respecto al total de instancias clasificadas como positivas. Se calcula como:

$$Presición = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

4. **Sensibilidad o Tasa de Verdaderos Positivos (Recall):** La proporción de verdaderos positivos respecto al total de instancias positivas en los datos reales. Se calcula como:

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

F1-Score: La media armónica de Precisión y Recall. Es útil cuando se busca un equilibrio entre ambas métricas.

$$F1 = 2 \times \frac{Presición \times Recall}{Presición + Recall}$$

5. **AUC (Área bajo la curva ROC):** Representa la capacidad del modelo para discriminar entre las clases. Cuanto más cercano sea el AUC a 1, mejor será el rendimiento del modelo. Si un modelo tiene un AUC de 0.5 se dice que el modelo tiene una capacidad similar al de un modelo aleatorio para diferenciar a las clases.

6. **Curva Lift**: La curva Lift compara el rendimiento del modelo con el de un modelo aleatorio. Muestra cómo el modelo mejora las predicciones en relación con una clasificación aleatoria.

$$Lift(k) = \frac{\text{Proporción acumulada de positivos reales hasta el k} - \text{ésimo percentil}}{\text{Proporción acumulada de positivos predichos hasta el k} - \text{ésimo percentil}}$$

La curva de Lift se traza para diferentes percentiles de probabilidad de predicción, proporcionando información sobre la mejora relativa en la precisión a medida que se avanza en la clasificación.

6.5. Interpretabilidad de Modelos

La interpretabilidad de modelos se refiere a la capacidad de comprender y explicar cómo funciona un modelo de aprendizaje automático, permitiendo que tanto expertos en el dominio como usuarios no expertos comprendan las decisiones o predicciones que realiza. Esta comprensión es esencial para ganar confianza en la aplicación práctica de los modelos y facilitar su adopción en entornos críticos.

En el caso específico de modelos basados en árboles de decisión, como XGBoost o Random Forest, se puede utilizar la técnica de SHAP values para analizar la importancia de las variables. Este método proporciona una explicación detallada y personalizada de la contribución de cada característica para una predicción específica. Estos valores asignan un peso a cada característica de entrada considerando cómo contribuye al valor de la predicción en comparación con su contribución promedio. Este nivel de detalle es especialmente valioso en modelos complejos, en donde la cantidad de variables es significante y no se puede analizar la interacción entre ellas de manera sencilla. En el caso de un modelo de clasificación binaria, con SHAP se puede asignar un ranking de importancia de variables y también ver cómo el valor de una variable indica qué tan positiva o negativamente contribuye a la probabilidad de pertenecer a una clase específica. Estos insights ayudan a comprender mejor la lógica interna del modelo y a tomar decisiones informadas basadas en la influencia de cada variable en las predicciones (Lundberg, 2017).

Capítulo 7: Metodología

7.1. CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es una metodología estructurada para el desarrollo de proyectos de minería de datos. Incluye descripciones de las fases y tareas necesarias en cada etapa, facilitando así el cumplimiento sistemático de los objetivos del proyecto con un enfoque efectivo y ordenado en todas las etapas. La justificación del uso de esta metodología radica en su capacidad para integrar la comprensión del negocio y de los datos en la construcción del modelo, también permite ajustar y mejorar el modelo de manera iterativa, lo que es de vital importancia para la consecución de los objetivos del proyecto (Wirth, 2000).

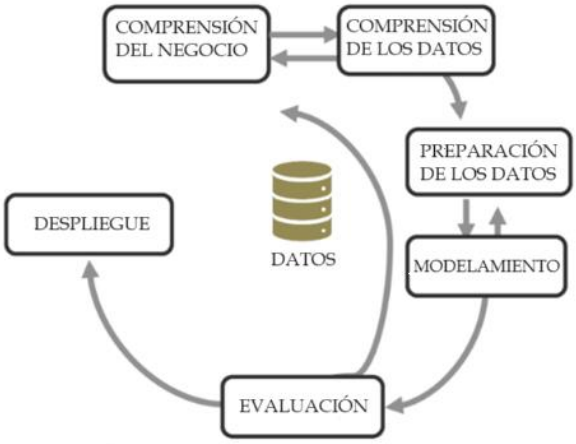

Ilustración 5: Diagrama de metodología CRISP-DM

Como se indica en la ilustración 4, la metodología CRISP-DM se divide en seis fases principales que se pueden describir de manera breve de la siguiente manera:

- **Comprensión del negocio:** En esta fase, se establecen los objetivos del proyecto desde una perspectiva comercial y se traducen en metas de minería de datos. También se evalúan los recursos disponibles y se define el éxito del proyecto.

- **Comprensión de los datos:** Se recopilan y exploran los datos disponibles para comprender su naturaleza. Esto incluye la identificación de posibles problemas de calidad de datos y la selección de las variables más relevantes para el análisis.

- **Preparación de los datos:** En esta fase, se realizan tareas de limpieza, transformación y procesamiento de datos para prepararlos para el modelado. Esto implica la selección de variables, manejo de valores atípicos, imputación de valores faltantes, etc.

- **Modelamiento:** Aquí es donde se seleccionan y aplican técnicas de modelado, se divide el conjunto de datos en entrenamiento y testeo, se selecciona modelos de aprendizaje automático, para construir y entrenar modelos con los datos preparados.

- **Evaluación:** Se evalúan y comparan los modelos construidos para seleccionar el mejor en términos de su capacidad para cumplir con los objetivos del proyecto. Esto implica el uso de métricas de rendimiento específicas, realizar ajustes en el modelo si es necesario y determinar la eficacia del modelo en relación con los objetivos del negocio.

- **Despliegue o Implementación:** Los modelos seleccionados se implementan en el entorno operativo y se integran en los procesos comerciales existentes. También se establece un plan de monitoreo y mantenimiento para asegurar el rendimiento continuo.

7.2. Herramientas

Para poder trabajar la metodología CRISP-DM se cuentan distintas herramientas e insumos. Como entorno de desarrollo, se hará uso de Jupyter Notebooks, esta herramienta proporciona un ambiente interactivo que favorece la experimentación y permite la documentación en tiempo real. Aunque menos estructurado en comparación con otros entornos, su naturaleza interactiva se alinea con las necesidades del proyecto.

Como lenguaje de programación se utilizará Python, se justifica debido a la alta disponibilidad de bibliotecas de ciencia de datos como Pandas, NumPy y Scikit-Learn. Su amplio uso en la comunidad de ciencia de datos, sumado al acceso a documentación actualizada, lo convierten en una elección ideal para realizar análisis exploratorio y modelado. Se incorpora la herramienta Pycaret, que es un framework de AutoML (He, 2021), que permite probar distintos modelos de clasificación con pocas líneas de código. Además, se emplearán herramientas de visualización como Matplotlib y Seaborn, las cuales ofrecen la posibilidad de comunicar efectivamente los resultados y contribuir a la comprensión de patrones y tendencias.

En cuanto a los recursos e insumos de datos, con el fin de reducir significativamente la necesidad de recolección de datos, se aprovecharán las bases de datos de registros de uso de la red almacenadas en Amazon Athena, que es un servicio de consulta interactiva de ambiente cloud que permite hacer procesos extracción, transformación y carga de datos. Ofrece la ventaja de reducir significativamente los tiempos de consulta de datos, aprovechando la eficiencia de la plataforma y del lenguaje SQL para la gestión y almacenamiento de grandes conjuntos de información.

Capítulo 8: Desarrollo

8.1. Comprensión del Negocio

Se llevaron a cabo reuniones estratégicas con profesionales del equipo de Analytics en experiencia de clientes, con el objetivo de comprender a fondo los aspectos asociados a los problemas de red y su vinculación con la fuga de clientes, definiendo así el objetivo del proyecto desde la perspectiva comercial. La meta es que la compañía pueda identificar a aquellos clientes que recientemente optaron por cambiarse a otra compañía debido a inconvenientes en la red, los cuales generaron una mala experiencia de usuario del servicio de telefonía móvil. Es crucial que este motivo de fuga sea explicado por variables que describan la experiencia de uso, permitiendo establecer métricas que

faciliten el seguimiento y prevenir la ocurrencia de problemas similares en el futuro. El objetivo final del negocio es disminuir la cantidad de fugas por motivos atribuibles a la red.

La empresa cuenta con información sobre cuántos y qué clientes se portan a otras compañías cada mes, se tiene un panorama general de la fuga, pero no se conoce la causa específica en caso. Para abordar este aspecto, se llevan a cabo encuestas de portabilidad mensuales dirigidas a una muestra de aproximadamente 600 clientes de telefonía móvil que han abandonado la compañía recientemente para optar por otro proveedor. De este grupo, en promedio el 90% pertenecen a la modalidad de pospago. Estos datos son útiles como etiquetas para un modelo de aprendizaje supervisado.

Los proveedores de telefonía móvil recopilan una amplia gama de datos relacionados con la experiencia del usuario en su red. Esto incluye información sobre llamadas y uso de datos móviles, así como datos de ubicación, señal y velocidad de internet móvil. Además, se registra detalles sobre los dispositivos utilizados y eventos específicos de la red como por el ejemplo el cambio de tecnología 4G a 3G. Es posible acceder a esta información para los clientes encuestados y generar datos de entrenamiento para el modelo y estudiar cómo se relaciona su experiencia de usuario con su motivo de fuga.

Desde el punto de vista del problema de ciencia de datos, el criterio de éxito del modelo es lograr discriminar las clases, lo que se discutirá en los resultados.

8.2. Comprensión de los Datos

Se recopilan los registros de encuestas de portabilidad con el propósito de realizar un análisis exploratorio. Este proceso implica una revisión detallada de las respuestas proporcionadas por clientes que cambiaron de una compañía de telefonía móvil a otra y que fueron clientes pospago. Durante este análisis, se identifican posibles problemas de calidad de datos y se seleccionan las variables relevantes.

Tabla 1: Encuesta de Portabilidad, conteo de registros y porcentaje de clase positiva.

Periodo	2023-01	2023-02	2023-03	2023-04	2023-05	2023-06	2023-07	2023-08	2023-09	2023-10
N° Encuestados	513	526	538	528	531	579	571	573	569	571
% Motivo Red	38.4%	32.2%	28.2%	28.1%	29.2%	28.45%	25.5%	29.3%	26.5%	27.7%

En la tabla 1 se muestra un resumen de los datos, se presenta el conteo de registros por periodo y el porcentaje de clientes que declararon fugarse de la compañía por problemas de red. Los datos se han recopilado a lo largo de varios meses, desde enero de 2023 hasta octubre de 2023. Se tienen en promedio 550 encuestas mensuales y la prevalencia tiene una media de 30%.

Las métricas estándar pueden ser engañosas en situaciones de desbalance. Aunque un modelo pueda tener alta precisión en la clase mayoritaria, puede fallar en identificar la clase minoritaria. En casos de desbalance moderado, métricas como *Recall, AUC y Lift* son más fiables, menos influenciadas por la distribución de clases. Incluso en una proporción de 30/70, se presentan desafíos para identificar la clase minoritaria.

Dado que estos registros contienen información sensible de los clientes, antes de la carga a Athena se implementa la práctica de anonimización, que implica desvincular la información directa que podría identificar a los clientes, asegurando que los datos utilizados para análisis y procesamiento cumplan con los estándares de privacidad y reducir el riesgo de exposición de información confidencial en caso de una violación de seguridad.

Los registros históricos de uso de la red se encuentran en Athena, estos se generan para cada cliente activo en el periodo. Se encuentran agrupados en tablas que las agrupan por tipo de medición. Algunos de los parámetros o métricas más relevantes que describen la experiencia del usuario en el servicio de telefonía móvil son:

- **Cobertura:** Intensidad, potencia y calidad de la señal 3G y/o 4G que recibe su dispositivo.
- **Tráfico Internet:** Tráfico subida/bajada, tiempo activo/inactivo, cantidad de celdas, cambios de tecnología. Estas se encuentran segmentadas por tecnologías 2G, 3G y 4G, además del total.
- **Trafico Voz:** Cantidad de minutos recibidos/emitidos, cantidad de llamadas recibidas/emitidas, cantidad de minutos clientes/no clientes, cantidad de llamadas clientes/no clientes, compañía favorita recibida/emitida. Estas se encuentran segmentadas por tecnologías 3G y 4G, además del total.
- **Antenas:** Celda favorita, tipo de celda favorita, zona de celda favorita, tráfico de celda favorita. Estas variables se encuentran segmentadas por tecnología 3G y 4G, además del total también están segmentadas por día/noche.
- **Datos Comerciales:** Antigüedad, cargo fijo del servicio, cantidad de líneas asociadas.

Todas las variables anteriormente descritas están disponibles en tablas relacionales en donde se estructuran de manera diaria para cada cliente. Algunas poseen una tabla hermana en donde las variables están agrupadas de manera semanal o mensual, lo que permite el cálculo de medias y desviaciones estándar a lo largo de los periodos.

Un problema en la calidad de los datos es que la tabla que guarda las variables de tráfico en internet solo está disponible desde abril del 2023, lo que implica que la cantidad de datos disponibles para el entrenamiento y evaluación del modelo se encuentra acotada temporalmente. Esta limitación temporal podría afectar la capacidad del modelo para aprender patrones, ya que la información previa a abril de 2023 no está disponible en la tabla mencionada. Otro problema es que no existen registros de estas métricas para la tecnología 5G. Desde un punto de vista estadístico contar con una muestra más amplia y con variables más representativas permite que el modelo logre capturar de mejor manera las características de la población, esto significa obtener modelos más robustos y precisos en comparación a cuando se tienen pocos datos.

8.3. Preparación de los datos

Para construir el tablón de datos de entrenamiento, se decide agrupar la información de las variables de redes para el mes anterior a la fuga del cliente y dejar como variable dependiente el motivo de fuga. Este enfoque temporalmente cercano captura las condiciones de red más relevantes y próximas a la decisión de un cliente de cambiar de proveedor.

Para implementar esta estrategia se utiliza el lenguaje SQL para construir una consulta en Athena, se inicia obteniendo la fecha y el motivo de fuga a través de la tabla de encuestas de portabilidad que se ha subido al sistema. Posteriormente, se selecciona el mes anterior a esta fecha como el período de interés para el cruce de datos. Se construye una Query que agrupa la información de cada una de las tablas mencionadas anteriormente de manera mensual, en la agrupación según la naturaleza de la variable se hacen distintas operaciones, como por ejemplo recuentos, sumas, promedios y desviaciones estándar. Después de cruzar el resultado es cada etiqueta de motivo de fuga junto a las variables descriptivas de uso de red para el mes anterior a la fuga del cliente. Se describen los pasos generales en la ilustración 6.

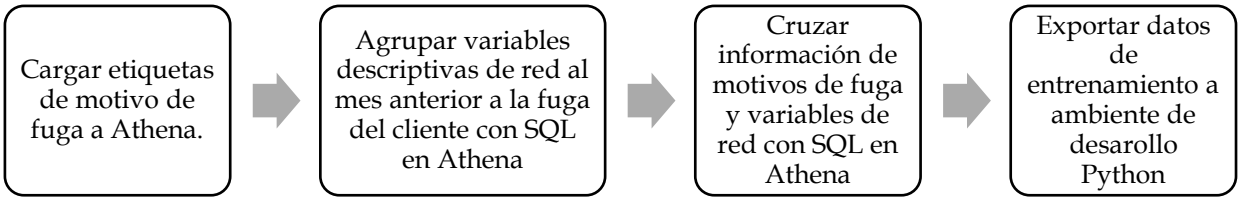

Ilustración 6: Proceso de cruce de datos de red y encuestas.

Una vez que se ha obtenido y estructurado el conjunto de datos, se procede a cargarlo en un entorno de trabajo de Jupyter Notebook. En este entorno y con la ayuda de librerías de python, se lleva a cabo una fase de limpieza y transformación de datos para abordar posibles problemas de calidad, inconsistencias o formato. Durante la limpieza de datos, se realiza el manejo de valores faltantes, la corrección de errores tipográfico e identificación de posibles outliers. En este caso cuando hay datos faltantes estos no son solo para una variable, sino que para todas las variables que tienen la misma tabla de origen, entonces no se puede hacer imputación de datos, solo se deben eliminar las filas ya que una gran cantidad de columnas poseen datos nulos. También se eliminan variables que presentan multicolinealidad para evitar representar la misma información en dos variables diferentes.

Como se indicó anteriormente, la tabla de variables de tráfico de internet posee información desde abril de 2023, por lo que se tienen que eliminar todos los registros anteriores a ese periodo. En la tabla 2 se puede ver la distribución de datos luego de la limpieza de registros, cabe destacar que después de la limpieza la prevalencia se mantiene cercana al número original en 29.3%.

Tabla 2: Distribución de registros después de limpieza de datos.

Periodo	2023-04	2023-05	2023-06	2023-07	2023-08	2023-09	2023-10	2023-10	Total
Registros	456	466	498	502	473	446	464	464	3305
% Motivo Red	37.4%	31.2%	30.2%	27.1%	30.2%	27.7%	24.4%	31.2%	29.3%

En el proceso de Feature Engineering se han creado 35 varias nuevas características a partir de las variables existentes en el conjunto de datos. Estas nuevas características buscan capturar relaciones, variabilidad y proporciones que podrían ser relevantes para el análisis. Se utilizan interacciones entre variables para calcular ratios, diferencias, sumas, multiplicaciones, comparación entre tecnologías, etc. Las interacciones tienen la finalidad de capturar la relación entre la cobertura, tráfico y conexión a antenas de los clientes, un ejemplo de variable construida que es relevante posteriormente es el score general de cobertura que se calcula ponderando el score 3G y 4G (variables preexistentes) y por el tiempo de uso y la cantidad de celdas por tecnología a las que se conecta el cliente a lo largo del mes.

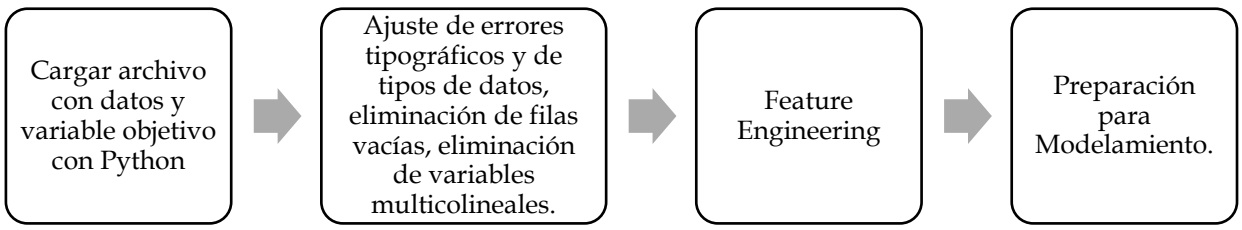

Ilustración 7: Proceso de limpieza, transformación y creación de nuevas variables.

El resultado final es un set de datos de entrenamiento de 256 variables junto a la variable objetivo y 3305 registros. Al evento que se quiere predecir, es decir si una fuga fue por problemas de red, se le llama clase positiva.

8.4. Modelamiento

Se empleó la Validación Cruzada Estratificada con 5 pliegues (folds) para dividir los datos en sets de entrenamiento y validación. Se utiliza este enfoque para abordar el desbalance de clases en el conjunto de datos, asegurando que cada clase estuviera representada de manera equitativa en todos los pliegues y asegurar que la evaluación del modelo no dependa de la elección al azar de una sola división de los datos, esto se hace evaluando el rendimiento del modelo posteriormente en cada pliegue, si varía mucho el rendimiento en cada uno significa que el modelo no está capturando tendencias generales de los datos, sino que solamente se está ajustando a los datos específicos con que se entrena en cada ciclo. En la ilustración 8 se índica como se separa el set de datos, en cada iteración se ajusta el modelo a los datos de entrenamiento y se crean predicciones con el set de validación para posteriormente evaluar el rendimiento con las métricas que se estimen convenientes, además se calcula el promedio de estas métricas.

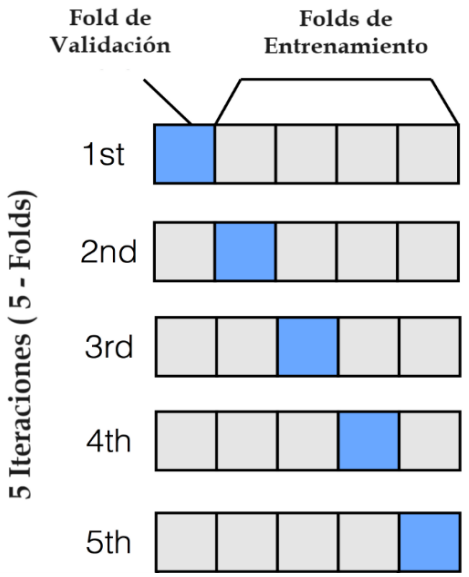

Ilustración 8: División de datos con Validación cruzada.

Esta estrategia ya está implementada en la librería PyCaret, por lo que se pueden ejecutar con pocas líneas de código a través de Jupyter Notebooks. La elección de PyCaret como framework de AutoML en Python se fundamenta en su capacidad para agilizar el proceso de comparación entre diversos modelos, la selección eficiente de hiperparámetros y la evaluación rápida del rendimiento de cada modelo.

La función "Comparar Modelos" de PyCaret ofrece una visión general y comparativa de múltiples modelos de aprendizaje automático en función de diversas métricas de evaluación. Esta herramienta permite identificar rápidamente los modelos más prometedores para el conjunto de datos en cuestión, brindando información sobre su rendimiento relativo. Para ello se programa entrenar los modelos de clasificación binaria que se describen en el marco conceptual y utilizar las estrategias de división de datos descrita para separar en entrenamiento y testeo.

8.5. Evaluación

Considerando el desbalance de clases en el conjunto, para identificar en primer lugar qué modelo es mejor para discriminar las clases se utiliza la métrica Área bajo la Curva ROC (AUC). Se utiliza la Validación Cruzada Estratificada con 5 folds para dividir los datos, se obtiene la métrica AUC en cada set de validación y se calcula el promedio entre ellos, este se transforma en una medida general del rendimiento, mitigando la sensibilidad a divisiones específicas de los datos. El modelo que mejor logra discriminar las clases a lo largo del set de datos es el que tiene mayor AUC promedio en validación. Un AUC cercano a 1 indica que el modelo tiene una sólida capacidad para distinguir entre clases. Cuando el AUC se sitúa en un rango moderado, por ejemplo, entre 0.5 y 1, se sugiere que el modelo tiene cierta capacidad para distinguir entre clases, aunque no de manera perfecta.

Esta estrategia se repite para distintas iteraciones, en donde se prueban estrategias como normalización y transformación de variables, creación de nuevas variables, eliminación de variables correlacionadas, eliminación de outliers, creación de datos artificiales de la clase minoritaria para equilibrar las clases. La función “Setup” de pycaret permite aplicar estos métodos al set de entrenamiento solo indicando su activación.

El modelo seleccionado con mejor rendimiento pasa a una segunda etapa de evaluación. En donde se compara el AUC promedio de entrenamiento y de validación para estudiar si existe el fenómeno de sobreajuste, en el caso de que exista, se estudian que hiperparámetros del modelo se pueden ajustar para prevenirlo y se vuelve a iterar con el entrenamiento del modelo.

Para estudiar cómo se comporta el modelo con otras métricas se genera una división simple aleatoria de los datos de 80% de entrenamiento y 20% de testeo y se entrena con los mismos hiperparámetros. Se analiza la curva Lift del modelo para obtener una representación gráfica de cómo el modelo clasifica las instancias positivas en comparación con un modelo aleatorio.

Después de la identificación de un modelo prometedor, se inicia un análisis detallado de métricas clave, como Precisión, Exactitud, Recall, F1 y la matriz de confusión. Es esencial señalar que la elección del umbral de clasificación tiene un impacto significativo en estas métricas. Se busca un umbral de decisión coherente que prediga una prevalencia similar a la observada en los datos de entrenamiento, manteniendo una Precisión y Recall coherentes. Se propone un análisis gráfico que representa estas métricas en función del umbral de clasificación. Esto proporciona una visión más completa y detallada del rendimiento del modelo, permitiendo una elección informada del umbral óptimo para crear nuevas predicciones de clases.

Para el análisis de interpretabilidad del modelo, en el caso de que el de mejor rendimiento sea un basado en arboles de decisión se propone utilizar SHAP Values, en otro caso se puede utilizar importancia de variables estándar del modelo.

Se resume el proceso de evaluación en la siguiente ilustración:

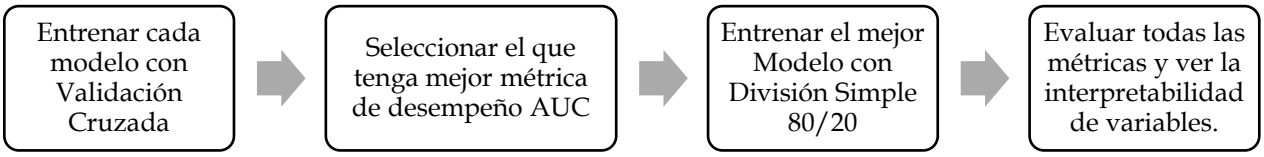

Ilustración 9: Descripción del proceso de Evaluación

El seguimiento de este desarrollo permite obtener el modelo con el mejor rendimiento posible bajo las condiciones y recursos existentes, el cumplimiento de los objetivos de minería de datos y en consecuencia de los objetivos del negocio dependen de los resultados.

8.6. Despliegue

La fase de despliegue e implementación del modelo seleccionado se inicia con el desarrollo del código de implementación. Esta vez se deja de utilizar Jupyter Notebooks y se comienzan a utilizar Scripts de Python en una carpeta estructurada pensada en el despliegue de un proyecto de ciencia de datos, que depende de las políticas de desarrollo de la compañía. Una vez que el código está listo, se procede con la integración efectiva de los modelos en el sistema de servicios de computación y almacenamiento en la nube existente. Esto implica asegurar una conexión adecuada con bases de datos, sistemas de gestión u otros componentes esenciales para el flujo operativo de la

organización, se considera la escalabilidad del sistema para adaptarse a volúmenes crecientes de datos para crear predicciones para un grupo grande de clientes.

Se crea la documentación que describe cómo acceder a los modelos, interpretar predicciones y otros aspectos relevantes. Se establece un plan de monitoreo continuo para evaluar el rendimiento del modelo en producción, que incluye la supervisión de métricas clave y la aplicación de actualizaciones a la obtención de nuevas encuestas de portabilidad.

Capítulo 9: Resultados

9.1. Benchmark de Modelos - AUC

A continuación, se presentan los principales resultados y hallazgos derivados del desarrollo del proyecto.

Tabla 3: AUC promedio para distintos modelos utilizando validación cruzada con 5 folds.

	Promedio AUC en 5 folds	
Modelo	Validation	Train
CatBoost Classifier	0.67	0.99
Gradient Boosting Classifier	0.66	0.93
Extra Trees Classifier	0.66	1
Random Forest Classifier	0.65	1
Linear Discriminant Analysis	0.64	0.77
Extreme Gradient Boosting	0.62	1
Decision Tree Classifier	0.55	1
Naive Bayes	0.52	0.53
K Neighbors Classifier	0.51	0.77
Logistic Regression	0.50	0.54
Dummy Classifier	0.50	0.50

Los resultados del benchmark de modelos realizado en la primera etapa de modelamiento y evaluación revelan que el algoritmo CatBoost Classifier presenta el mejor desempeño en la métrica de AUC promedio (área bajo la curva ROC) en validación cruzada, alcanzando un valor de 0.67 según la Tabla 3. Esto implica una alta capacidad del modelo CatBoost para distinguir entre clientes que efectivamente abandonaron Entel por insatisfacción de red, de aquellos que no tienen un alto riesgo de fuga técnica.

Sin embargo, al comparar su AUC de entrenamiento -que llega a 0.99- se observa una divergencia importante respecto al valor obtenido en validación. Esta diferencia manifiesta un sobreajuste (overfitting) de CatBoost a los datos de entrenamiento.

El overfitting o sobreajuste de un modelo ocurre cuando este se ajusta demasiado a los datos específicos con los que fue entrenado, perdiendo capacidad de generalizar bien ante nuevos datos. Se puede detectar comparando métricas de performance del modelo entre el conjunto de entrenamiento y uno de validación/prueba. Si la métrica (por ejemplo, Accuracy, AUC, etc.) es significativamente superior en training, indica overfitting.

Esto ocurre típicamente en modelos muy complejos, con muchos hiperparámetros o que se entrenan por demasiadas iteraciones. Memorizan patrones específicos que no se mantienen en datos nuevos. El overfitting es problemático porque genera una falsa sensación de precisión durante el entrenamiento, la cual no se materializa al hacer inferencias con nuevos datos. Se pierde capacidad de generalización, lo cual degrada seriamente la utilidad en el mundo real. Existen técnicas como regularización, early stopping y obtención de más datos que ayudan a mitigar este problema y mejorar la generalización.

Pese a aquello, el desempeño en validación cruzada demuestra que aún ante este overfitting, el modelo logra generalizar bien sobre datos no observados previamente y supera en performance a los demás algoritmos evaluados.

9.2. Modelo CatBoost Optimizado - AUC

Se realizó un ajuste de hiperparámetros del modelo CatBoost para abordar el sobreajuste observado. Se redujo el número de iteraciones (cantidad de árboles de decisión ensamblados) de 1000 a 150 para disminuir la complejidad del modelo, este valor indica cuantos arboles se usan para ensamblar el modelo. Se utilizó la regularización L2 en los nodos hoja para disminuir la profundidad de los árboles mediante una penalización más fuerte.

Tabla 4: Modelo CatBoost con Ajuste de Hiperparámetros y validación cruzada con 5 folds.

	Promedio AUC en 5 folds	
Modelo	Validation	Train
CatBoost Classifier – Hiperparámetros optimizados	0.71	0.77

Tabla 5: Modelo CatBoost con Ajuste de Hiperparámetros y division simple train/test.

	AUC – División Random 80/20	
Modelo	Test	Train
CatBoost Classifier – Hiperparámetros optimizados	0.68	0.76

Al implementar el modelo con ambas estrategias de división de datos, se puede apreciar en la tabla 4 y 5, que al disminuir el sobreajuste el modelo gana capacidad de generalizar, aumentando su capacidad de distinguir entre las clases.

9.3. Selección de un umbral de decisión

Para estudiar otras métricas, se hace una nueva división aleatoria 80% train y 20% test y se reentrena el modelo.

Ilustración 10: Métricas de desempeño en función del umbral de decisión.

El umbral de decisión (*threshold*) utilizado para convertir los scores de predicción del modelo en clases binarias (0 o 1) posee una influencia crucial tanto en la precisión final como en el ratio de casos positivos detectados. Si bien por defecto suele utilizarse un valor de 0.5, el análisis realizado muestra que este valor no maximiza necesariamente el desempeño en el conjunto de datos, ya que para este umbral el modelo no es capaz de detectar ni un 20% de la clase positiva, y esto se agrava por el grado de desbalance entre las clases minoritarias (clientes que efectivamente abandonan por motivos técnicos) y la clase mayoritaria (clientes que desertan por otros motivos).

En la ilustración 10 se analizan los resultados en el set de testeo para distintos umbrales de decisión, se gráfica el valor de distintas métricas en el eje Y, estas métricas se entienden como porcentajes.

Se puede ver que mientras más alto es el umbral de decisión, menos casos son catalogados como positivos.

Mediante esta representación del performance del modelo para distintos umbrales, pudo determinarse un valor alternativo de 0.32 como *threshold*, este valor aumenta la tasa de positivos en la predicción del 10% al 30%, que es el valor real en los datos. Este es el punto donde la curva de *precisión* y de *Recall* alcanzan su valor en conjunto más alto, convirtiéndose en el punto en donde se mantiene una tasa de positivos controlada sin descuidar la precisión general del modelo, esto es crítico para uso eficiente de recursos en el caso que se necesiten detectar la cantidad de clientes que se fugan por motivos técnicos sin sobreestimar su cantidad.

9.4. Métricas de Desempeño

Las métricas de desempeño para el modelo con umbral de 0.32 se presentan como resultados en la tabla 6.

Tabla 6: Métricas de desempeño para umbral 0.32

	Accuracy	Recall	Prec.	F1	Positivos
Test	0.69	0.46	0.46	0.46	0.29
Train	0.74	0.58	0.53	0.55	0.31

El modelo tiene una exactitud (*Accuracy*) del 69%, lo cual indica la proporción de predicciones correctas en general. Sin embargo, la exactitud por sí sola puede no ser suficiente si hay un desequilibrio en las clases. El valor AUC es 0.68, lo cual sugiere que el modelo tiene un rendimiento razonablemente bueno al clasificar entre clases positivas y negativas. El *Recall* dice que el modelo captura el 46% de todas las instancias positivas. Es importante en el caso de que se necesiten capturar todas las variables positivas. La precisión indica que el 46% de las instancias predichas como positivas son realmente positivas. Es importante si la minimización de falsos positivos es una prioridad.

Tabla 7: Matriz de confusión para umbral de 0.32

	Predicted Negative	Predicted Positive
Actual Negative	374	100
Actual Positive	102	85

En la tabla 7 podemos apreciar la matriz de confusión para los datos de testeo del modelo, podemos ver que no es un modelo perfecto y comete varios errores, pero si marca una tendencia general, en

ese sentido al crear una predicción especifica no se tendrá mucha certeza si es correcta o no, pero al hacerlo por grupos de clientes y analizar la vista general si podremos diferenciar tendencias.

9.5. Curva Lift

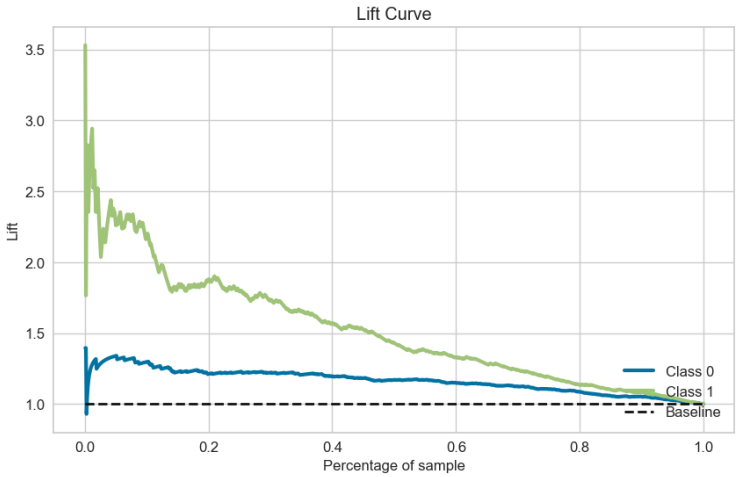

Ilustración 11: Curva Lift de modelo Catboost - train/test 80/20 aleatorio

La curva Lift es una forma muy útil de evaluar y visualizar el rendimiento de un modelo de clasificación, específicamente para problemas como la predicción de fuga de clientes. Básicamente la curva de Lift muestra cuántas veces mejor (en términos de encontrar clientes fugados reales por motivos técnicos) es nuestro modelo comparado con un modelo aleatorio. En la ilustración 11 podemos ver que el modelo es capaz de identificar la clase positiva de buena manera.

Como se muestra en la ilustración 10, el problema es que es difícil determinar el umbral óptimo y pequeños cambios en el umbral pueden impactar significativamente la precisión. En un enfoque basado en propensión, en lugar de predecir clases explícitamente, se ordenan los ejemplos según la probabilidad de clase 1. Los ejemplos más propensos (es decir, con mayor probabilidad de clase 1) se pondrían arriba. Esto permite crear distintos grupos de ejemplos. Por ejemplo, se podría definir un grupo de "alto riesgo" con los 100 ejemplos más propensos, un grupo de "riesgo medio" con los siguientes 200, etc.

La principal ventaja de utilizar este enfoque es que no se requiere establecer un umbral, el cual puede ser difícil de afinar, permite crear segmentos naturales y diferentes grados de riesgo. El modelo no cambia, solo la forma de usar las probabilidades predichas. En este caso la curva Lift

refleja una concentración de casos reales de deserción por motivos técnicos de red entre los clientes que el modelo asigna dentro del top 10, 20, 30% de mayor probabilidad de fuga. Sin necesidad de umbral se aprovecha así la efectividad de priorización demostrada por el modelo dentro de la población con los scores de propensión más altos. Esto es de utilidad ya que se pueden filtrar y agrupar a los clientes con alta propensión y ver en que zonas geográficas se mueven y analizar por qué están teniendo una mala experiencia de usuario con la red de telefonía móvil de Entel.

9.6. Interpretabilidad de Variables

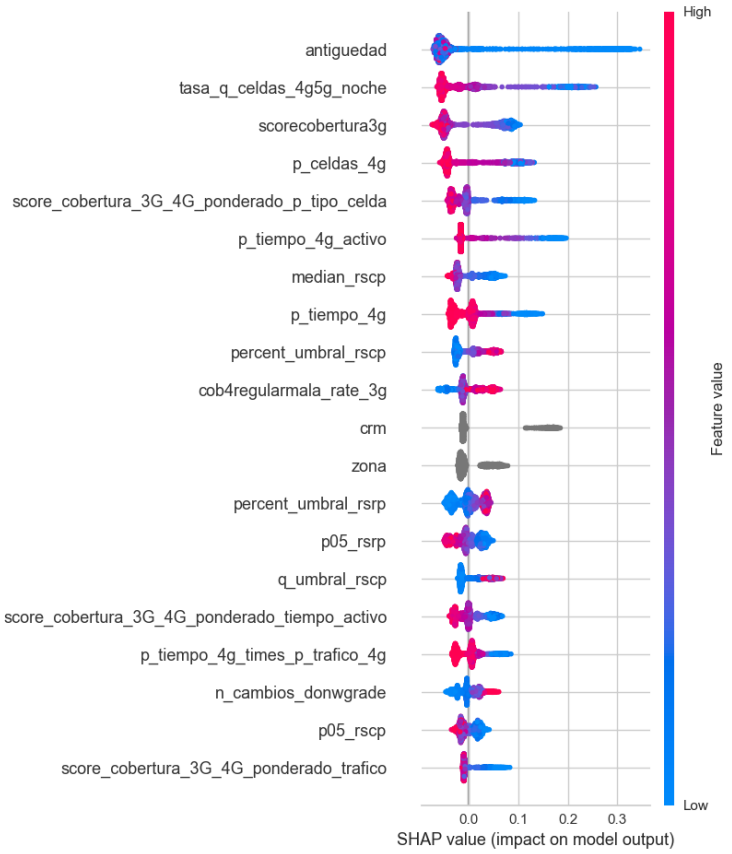

Ilustración 12: Análisis de Interpretabilidad con SHAP Values – Modelo CatBoost con Hiperparámetros Ajustados

Se utiliza el gráfico “Summary Plot” de la librería SHAP para hacer un análisis de interpretabilidad del modelo y entender que variables son relevantes para el problema. Para entender el gráfico, hay que entender que las variables más importantes se encuentran más arriba. El color del punto indica el valor de la característica para la instancia específica. Un punto rojo significa que esa característica tiene un valor bajo en ese registro. El gráfico también incluye una barra vertical en el centro que representa la predicción promedio del modelo para todas las instancias. Los puntos que están a la derecha contribuyen positivamente a la predicción promedio, mientras que los puntos a la

izquierda contribuyen negativamente. Se puedes ver cómo el valor de la característica se relaciona con el impacto de esa característica en la predicción.

El modelo muestra nos dice que las cinco variables más importantes son la antigüedad, el porcentaje de celdas 4G-5G del total en horario noche a la que se conecta el cliente, el score de la cobertura 3G que recibe, porcentaje de celdas 4G en comparación al total y el score de cobertura 3G y 4G ponderado por el porcentaje de celdas de cada tecnología al que se conecta el cliente. En el gráfico se puede analizar como el valor de cada variable aporta a que el cliente se fugue por motivo de red dependiendo de la ubicación y color de los puntos, se puede ver que clientes nuevos (puntos azules) son más propensos a portarse por problemas de red (están hacia la derecha).

Tal como se menciona, la red 3G tiene menor capacidad para manejar grandes volúmenes de tráfico de datos. Por lo tanto, en horarios de alto uso como la noche cuando más usuarios están conectados consumiendo video/música, es mucho más probable que un cliente 3G experimente lentitud, interrupciones de servicio y caídas de red por saturación.

Esto claramente genera una percepción de baja calidad e insatisfacción mucho mayor en clientes de redes 3G comparado con 4G más robusta. Ese es un motivo por el cual el modelo de detección de fuga identificó el porcentaje de uso 4G/5G en horario nocturno como un indicador clave para predecir el riesgo de abandono.

Se construyeron visualizaciones para entender a detalle como la distribución de estas variables generan diferencias en la predicción del modelo y como se compara con el dato real. Para esto se calcula el porcentaje de positivos en los datos reales del set de testeo contra el promedio de la propensión predicha por el modelo, que se entiende como el valor esperado o porcentaje de positivos predichos.

En las ilustraciones 13 y 14 se detalla el análisis de la antigüedad y del porcentaje de celdas 4G-5G al que se conecta el cliente en las noches y como afecta su distribución a los resultados del modelo.

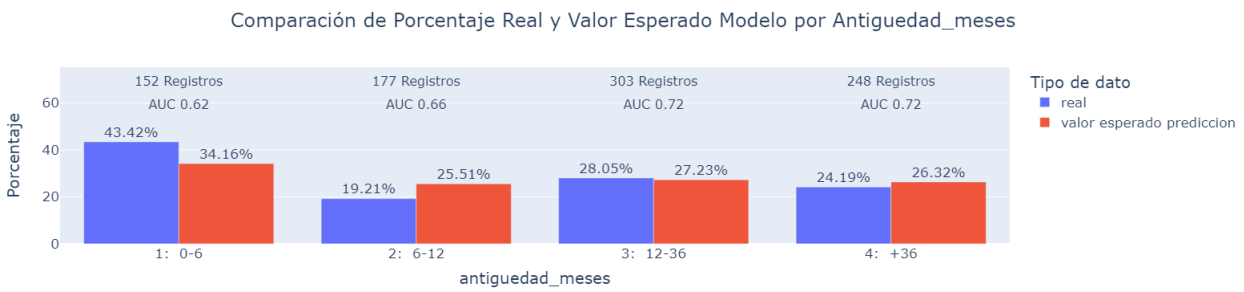

Ilustración 13: Análisis de predicción vs dato real en variable antigüedad

En esta visualización se ve que clientes que tienen antigüedad de 0 a 6 meses son los más propensos a portarse por un problema de red, mientras que los más antiguos son los menos propensos. El modelo es capaz de capturar esta tendencia, ya que la distribución de los datos reales es bien cercana a la de las predicciones. Esto se puede explicar por el simple hecho de que los clientes de mayor antigüedad ya se hubieran fugado si el servicio funcionara mal. Un cliente nuevo al experimentar un servicio insatisfactorio es probable que busque otra opción.

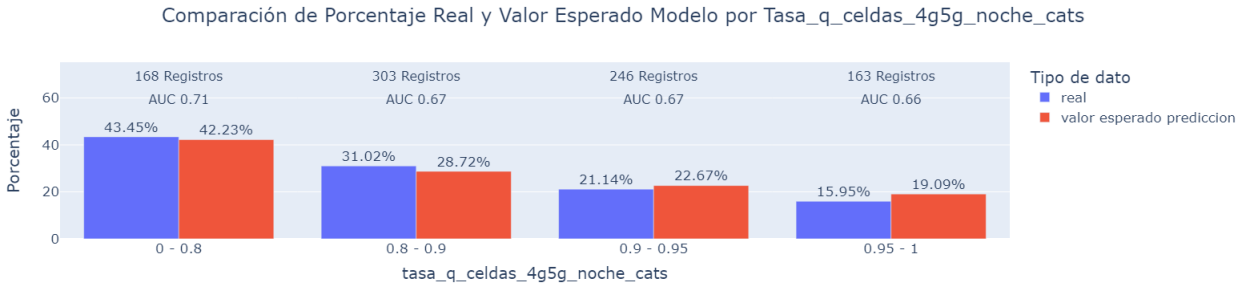

Ilustración 14: Análisis de predicción vs dato real en variable tasa de celdas 4G-5G noche.

Similar con el caso de la variable que indica el porcentaje de celdas 4G-5G a la que se conecta el cliente en la noche, se puede ver que un cliente que tiene acceso a redes 4G en la noche tiene menos probabilidad de que tenga un servicio insatisfactorio que lo lleve a portarse por un motivo de red, mientras que un cliente que utilice al menos un 20% de antenas 3G-2G en la noche aumenta la probabilidad significativamente. El tipo de tecnología que el cliente utiliza en la noche se convierte en una variable importante para describir la experiencia de uso del usuario, es normal que en el horario 19:00 – 23:59 las personas estén más desocupadas y el nivel de tráfico sea más alto. La tecnología 3G tiene menos capacidad de volumen de tráfico que la tecnología 4G, es probable que una antena 3G se sature antes y que esto pueda llevar a una velocidad de descarga lenta y tiempos de espera prolongados.

9.7. Análisis Geográfico

Con el objetivo de identificar la existencia de patrones geográficos vinculados a una mayor propensión de abandono de clientes por insatisfacción de red, se realizó una representación territorial de los scores promedio de fuga técnica pronosticados por el modelo CatBoost. En este caso se utilizó la base de clientes que se portaron en noviembre del 2023, alrededor de 50 mil en total.

La visualización se construyó georreferenciando a los clientes en base a las 3 celdas con su mayor tráfico histórico y asignándolos a hexágonos representativos de tamaño variable según el nivel de agregación deseado. Posteriormente se calculó la propensión media dentro de cada hexágono según los valores individuales de sus clientes constituyentes. Los polígonos resultantes se muestran en un mapa térmico, donde las tonalidades más oscuras indican zonas con las mayores probabilidades pronosticadas de salida por red.

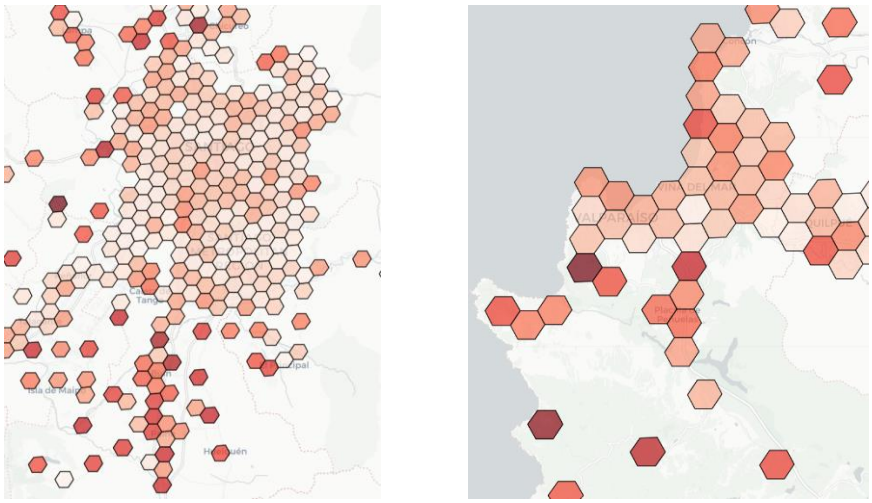

Ilustración 15: Representación geográfica de propensión de portabilidad por motivo de red en Región Metropolitana (izquierda) y zona urbana de Valparaíso (derecha).

Esta vista territorial permite identificar varias áreas geográficas donde se concentran clientes para los cuales el modelo predice una elevada propensión al abandono fundamentado en una experiencia negativa de red móvil. Los insights respecto a donde están situados estos usuarios más susceptibles de desertar por fallas de servicio representan insumos de alto valor para investigar particularidades técnicas del funcionamiento de la red en dichas localidades. Así se pueden diagnosticar problemas relativos de infraestructura o capacidad y focalizar acciones mitigatorias.

Capítulo 10: Conclusiones

El objetivo general era desarrollar un modelo de clasificación binaria para predecir si un cliente presenta alto riesgo de abandonar la compañía debido a una mala experiencia técnica en la red móvil. Los objetivos específicos incluían: procesar datos para identificar patrones predictivos de esta fuga técnica; diseñar nuevas variables explicativas a partir de los datos; entrenar y evaluar algoritmos para estimar la probabilidad individual de fuga técnica, e implementar el mejor modelo para uso de equipos técnicos y comerciales.

Luego del análisis, se logró cumplir estos objetivos mediante la aplicación de técnicas de ciencia de datos, obteniendo un modelo de Boosting CatBoost con capacidad adecuada para distinguir entre clientes propensos y no propensos a abandonar por fallas de red. Se cumplió satisfactoriamente con el desarrollo propuesto del modelo de detección de fuga técnica, logrando un clasificador de buen desempeño. Sobre el modelo en sí, se resalta la identificación de drivers relevantes asociados a las redes 3G, ratificando la importancia crítica que tiene para las telefónicas proporcionar un servicio de red consistente en las tecnologías que ofrece y que mantenga satisfechos a los usuarios.

El rendimiento presente del modelo de predicción de fuga técnica es adecuado teniendo en cuenta la escasez actual de datos positivos confirmados para el entrenamiento (solo unos miles). Sin embargo, incrementar sistemáticamente el número de observaciones de portabilidad etiquetadas según su causa específica representa una vía tentativa de mejora continua.

Más datos permitirían entrenar modelos sin riesgos de sobreajuste y detectar interacciones más complejas entre un amplio conjunto de predictores de red. Además, se dependería menos del porcentaje aleatorio de datos destinados a las muestras de validación.

Bajo la premisa de que Entel tiene una cultura de datos, se propone implementar un proceso para encuestar una muestra representativa de clientes que abandonaron el servicio, determinar cuáles lo hicieron explícitamente por insatisfacción técnica e incorporar dichos casos confirmados para reforzar cíclicamente el conjunto de entrenamiento junto a sus variables de contexto de red. De esta manera, se podrá aprovechar todo el potencial de algoritmos como LightGBM o CatBoost a medida que crezca la base histórica de fugas validadas y posiblemente mejorar la capacidad predictiva del modelo.

La ausencia de métricas de red 5G pese a la existencia de tráfico de clientes en esa tecnología implica algunas desventajas importantes. No contar con indicadores de desempeño específicos de

la red 5G implica una pérdida de información valiosa sobre la experiencia de estos usuarios. Al no monitorear la red de la misma forma que se hace con otras tecnologías, no es posible determinar en qué medida problemas técnicos en esta nueva generación de red avanzada pudieran estar motivando abandono o insatisfacción del servicio. Tampoco podemos evaluar si existen diferencias significativas en el comportamiento de la red 5G comparado contra la generación anterior 4G que pudieran estar impactando en la tasa de abandono técnico de usuarios 5G. Por ejemplo, identificar si este segmento muestra una fuga mayor por fallas de red dadas expectativas más altas sobre la tecnología.

Asimismo, el crecimiento esperado en la proporción de clientes que cursen parte importante de su tráfico en la nueva red 5G pone en riesgo la efectividad a futuro del modelo de detección de fuga técnica si no se incorporan predictores específicos asociados. En poco tiempo, es probable que un segmento importante de usuarios interactúe predominantemente con esta tecnología.

Se recomienda continuar enriqueciendo los datos de entrenamiento con nuevos casos confirmados de fuga técnica, evolucionar las variables predictivas para representar experiencia de usuarios 5G e incorporar otras fuentes como interacciones de centros de atención que aporten señales adicionales de insatisfacción latente con el servicio de conectividad móvil recibido. De esta forma se podrá aprovechar al máximo la analítica avanzada para convertirla en una ventaja competitiva frente a la amenaza de portabilidad.

El objetivo desde el punto de vista de la ciencia de datos es obtener un modelo que pueda distinguir entre las clases, siendo capaz de identificar con un margen de error acotado a los clientes que se fugaron por problemas de red. En contraste, el objetivo del negocio es disminuir la tasa de fuga por motivos atribuibles a la red. En ese sentido se identifican dos formas concretas en que el modelo predictivo puede contribuir a Entel:

Optimización de campañas comerciales de retención: El modelo permite priorizar descuentos y beneficios sobre la base de clientes con baja propensión de salida técnica, mediante los scores de riesgo individuales. Enfocar recursos en usuarios sin problemas evidentes de red, pero proclives al cambio ante mejores ofertas de precio.

Diagnóstico focalizado de la red: Los mapas de calor con propensiones medias de abandono técnico por zona geográfica facilitan la detección de áreas críticas. Se pueden dirigir recursos de ingeniería a un análisis en profundidad del funcionamiento real de la red móvil en dichas localidades para identificar fallas y solucionar los problemas específicos que degradan la experiencia del cliente.

De esta manera el modelo entrega inteligencia accionable tanto para mitigar fugas ya latentes (campañas) como para prevenir problemas emergentes de infraestructura que podrían desencadenar nuevas fugas técnicas (diagnóstico de red).

Bibliografía

Alloghani, M. A.-J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, 3-21.

América Economía. (23 de octubre de 2020). *Por qué el precio del internet móvil es tan bajo en Chile y tan alto en México.* https://www.americaeconomia.com/articulos/por-que-el-precio-del-internet-movil-es-tan-bajo-en-chile-y-tan-alto-en-mexico#:~:text=Adem%C3%A1s%20de%20la%20competencia%20u,medio%20de%20acceso%20a%20internet.

Entel S.A. (2020). *Todo Chile comunicado.* https://www.entel.cl/pdf/todo_chile_comunicado.pdf

Entel S.A. (2023). Memoria Integrada 2022. Santiago, Chile.

Freeman, E. A. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological modelling, 217(1-2)*, 48-58.

García, J. M. (2018). *Ciencia de datos. Técnicas Analíticas y Aprendizaje Estadístico.* Bogotá, Colombia.: Publicaciones Altaria, SL.

He, X. Z. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems, 212*.

Lundberg, S. M. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, 381-386.

Prokhorenkova, L. G. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Subtel. (2012). *Portabilidad Numérica*. https://www.portabilidadnumerica.cl/que-es-la-portabilidad-numerica/

Subtel. (17 de Octubre de 2018). *Primera radiografía de consumo de datos de los chilenos revela que se utilizan principalmente en streaming de video*. https://www.subtel.gob.cl/primera-radiografia-de-consumo-de-datos-de-los-chilenos-revela-que-se-utilizan-principalmente-en-streaming-de-video/

Subtel. (Septiembre de 2023). *Reporte Mensual de Portabilidad*. https://www.portabilidadnumerica.cl/wp-content/uploads/2023/11/Reporte_Portabilidad_2023_09_30_v1.pdf

Subtel. (2023). *Sector Telecomunicaciones Primer Trimestre 2023.* https://www.subtel.gob.cl/wp-content/uploads/2023/06/PPT_Series_MARZO_2023_V0.pdf

Vuk, M. &. (2006). ROC curve, lift chart and calibration plot. *Advances in methodology and Statistics, 3(1)*, 89-108.

Wirth, R. &. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, (Vol. 1, pp. 29-39).