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

DISEÑO DE UN MODELO DE ATRIBUCIÓN DE FUGA DE CLIENTES A RAZONES ASOCIADAS A FALLAS TÉCNICAS DE REDES EN UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CRISTIÁN ALEJANDRO TORRES MORENO

PROFESOR GUÍA:
ANDRÉS GORMAZ CANAVE

MIEMBROS DE LA COMISIÓN:
BLAS DUARTE ALLEUY
ELIECER PEÑA ANCAVIL

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Cristián Alejandro Torres Moreno
FECHA: 2024
PROF. GUÍA: Andrés Gormaz Canave

DISEÑO DE UN MODELO DE ATRIBUCIÓN DE FUGA DE CLIENTES A RAZONES ASOCIADAS A FALLAS TÉCNICAS DE REDES EN UNA EMPRESA DE TELECOMUNICACIONES

La industria de las telecomunicaciones ha estado en los últimos años sujeta a una demanda cada vez más exigente, así como a una competencia más alta y a una economía inestable. Este trabajo se realiza en Entel, una empresa que, en base a su estrategia enfocada en ofrecer una experiencia distintiva gracias a una infraestructura superior, ha logrado ser líder en la industria. Sin embargo, en el último año el nivel de fuga del segmento pospago fue mayor al año anterior, afectando a la vez los resultados y la imagen corporativa de Entel de ser líderes.

Desde Entel se desconocen los motivos reales de estas fugas. No obstante, se sabe mediante una encuesta enviada a un 1% de esta población que un 30% lo hace por motivos de mala satisfacción de red, lo cual representa un quiebre con la estrategia de la empresa. En ese sentido, en este trabajo se genera una solución analítica dirigida a conocer los niveles totales de fuga de clientes debido a una mala experiencia o fallas técnicas de redes. En particular, se genera un modelo de atribución de si una fuga se debe o no a motivos de red, el cual asigne esta atribución a las antenas (o POP) con el fin de apoyar el plan de priorización de inversión de mejoras a los POP del año siguiente y, con ello, poder disminuir la fuga. El rol asumido en el trabajo consiste en desarrollar la etapa llamada MVP dentro de lo que compone un proyecto de datos en la empresa, lo que deriva en cada uno de los resultados de este informe.

Para desarrollar el modelo se utiliza la metodología CRISP-DM, utilizada comúnmente y de forma transversal entre las industrias para proyectos de datos. Además, producto de estar en una etapa intermedia global, se hace necesario realizar un levantamiento de la situación actual del proyecto y los insumos disponibles de fases previas. Por último, para aumentar la confianza de adopción como producto luego de este trabajo, se realizan análisis de robustez del modelo en el tiempo y se documenta el trabajo en estándares de la empresa.

En el desarrollo se obtuvo un modelo XGBoost que fue capaz de minimizar el sobreajuste a la vez de maximizar su métrica AUC, en donde se descubrió que los POP con poca cobertura 4G tienden a tener una peor experiencia. Además, se obtuvo una estabilidad y consistencia en un horizonte de 3 meses en el comportamiento de las predicciones de atribución, continuidad de pop críticos y correlación de fuga con atributos relevantes para el negocio.

El proyecto cumple con los objetivos y aporta un estudio de la fuga desde un enfoque novedoso. Sin embargo, existe espacio de mejora en los resultados en relación con planificar de forma más rigurosa los tiempos y plazos de cada etapa, esto enfocado en realizar un mejor levantamiento de datos y en tener un mayor respaldo científico en las decisiones tomadas, teniendo un contacto y una validación constante y directa con el negocio y sus necesidades.

A todo lo que he sido,
soy y seré.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

1. Antecedentes generales

1.1 Descripción del entorno y la empresa

1.1.1 Ambiente competitivo

El ambiente competitivo en el cual se está inmerso es en la industria de las telecomunicaciones, la cual a su vez está compuesta por segmentos de servicios de internet fija, telefonía móvil, telefonía fija y televisión de pago [1]. Para efectos del proyecto el entorno competitivo de interés es el de los servicios de telefonía móvil (o voz) ofrecido a los clientes tipo persona, lo que comprende la navegación de estos (con un dispositivo y número asociado) a través de las redes móviles utilizando datos de internet móvil o realizando actividades de voz, esto mediante las tecnologías 2G, 3G, 4G y la recientemente incorporada 5G en el 2022.

Estas tecnologías han ido variando de manera creciente en su uso y demanda a nivel país con el tiempo a medida que se incorporan nuevas y van quedando atrás otras como la 2G y 3G, las cuales han sido progresivamente reemplazadas por las 4G y 5G, que presentan en su conjunto una mejor cobertura, así como una mayor velocidad de carga y descarga de datos desde y hacia la red, respectivamente. Una evolución del uso de estas tecnologías en el tiempo hasta diciembre del 2023 [2] se puede ver en el Anexo A.

En Chile, el año 2023 fue un periodo difícil para la industria, que al igual que otros sectores, ha debido enfrentar los efectos del estancamiento económico y un alza sostenida en sus costos producto de la inflación [3]. Además, en palabras del presidente del directorio de Entel, "Las empresas de telecomunicaciones en Chile están enfrentando importantes desafíos de rentabilidad", destacando aspectos tales como consumidores exigentes y alta competencia [4], estos aspectos han hecho que las estrategias de retención de clientes y la eficacia operativa enfocada en la calidad del servicio ofrecido sean cada vez más desafiantes.

Con todo esto, considerando la participación de mercado que presentan los distintos exponentes de este rubro, Entel sigue siendo la que lidera en conexiones de internet móvil sumando las tecnologías 3G, 4G y 5G, llegando a diciembre del 2023 a una participación del 34.6% [2]. Por otro lado, en lo que respecta a los abonados móviles (clientes suscritos a telefonía móvil), Entel también ha liderado la participación, alcanzando un 32% a diciembre del 2023 [2], lo cual son 0.9 puntos porcentuales menor respecto al mismo período del 2022, de esto se profundizará en la siguiente sección.

La evolución en los últimos años de participación tanto en conexiones como en abonados móviles de los distintos exponentes se pueden ver en las figuras (a) y (b) de la Ilustración 1, respectivamente. Por otro lado, el detalle de los porcentajes al cierre de 2022 y 2023 (de [2]) se puede ver en los Anexos B (a) y (b), respectivamente para conexiones y abonados.

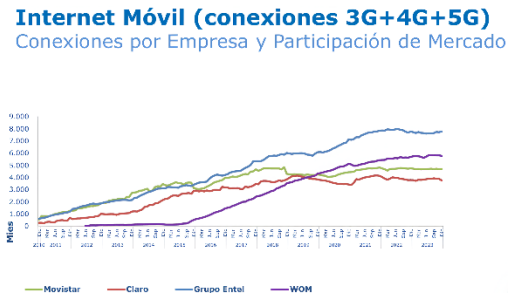


Figura (a): Participación por conexiones

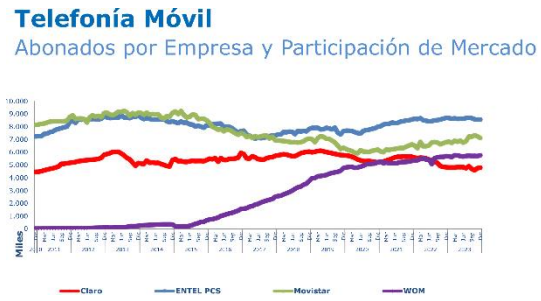


Figura (b): Participación por abonados

Ilustración 1: Evolución de participación en conexiones y abonados móviles (obtenido de [2])

1.1.2 Características de la empresa

Entel (Empresa Nacional de Telecomunicaciones) es una compañía dedicada a las tecnologías y telecomunicaciones, la cual cuenta con operaciones en Chile y Perú. Fue fundada en el año 1964, con el objetivo de asegurar la continuidad de las telecomunicaciones interurbanas y proveer conectividad de larga distancia internacional [5]. Entel ofrece servicios de conectividad móvil y fija, así como una amplia gama de servicios TI y digitales para los segmentos de personas (mercado business to costumer o B2C), empresas y grandes corporaciones (en su conjunto mercado business to business o B2B), y tanto en Chile como en Perú ofrece servicios mayoristas y de call center [3].

La administración de Entel Chile se estructura en segmentos de mercado que atienden las necesidades de los distintos clientes: Mercado personas, mercado empresas y mercado corporaciones. Todas las unidades operan bajo el liderazgo de la gerencia general [6]. En el Anexo C se puede apreciar un vistazo a la estructura del gobierno corporativo y las gerencias que rigen a Entel. Además, se puede apreciar de igual manera en el Anexo C la estructura interna dentro del área en donde se está realizando el proyecto, el cual se encuentra dentro de la subgerencia de Analytics y con el equipo de data scientists encargado de redes. En esa línea, también existe una colaboración estrecha con el equipo de Customer Value Management (o CVM), el cual en el proyecto tiene el rol de aportar una mirada más enfocada en comprender mejor la experiencia del cliente en relación a qué datos utilizar y a cómo validar la efectividad de la solución.

Entel ha centrado su estrategia de negocios acorde a los retos que afronta en la industria de las telecomunicaciones, sobrellevando estos desafíos a través de la experiencia distintiva, la fortaleza de marca y una solidez financiera que ha sido clave para capturar las nuevas oportunidades de la evolución tecnológica [3]. En esa línea, el liderazgo y retención de Entel en el mercado B2C en la industria ha estado apoyado por un alto reconocimiento de marca, una infraestructura superior y la experiencia multicanal que ofrece a sus clientes. Lo anterior, respalda lo declarado por Entel como su imagen corporativa, la cual se fundamenta en 3 pilares, siendo una de ellas tener “una gran señal” [3]. Resumiendo estas secciones, se presenta un análisis FODA de Entel en el Anexo D.

1.1.3 Impacto y abandono de abonados pospago

Los ingresos que suponen los servicios móviles para Entel representaron aproximadamente un 55% del total de los ingresos el año 2023. En particular, los ingresos asociados a los servicios móviles prestados al segmento pospago representan casi un 93%. Además, el servicio de telefonía móvil proporcionado a los clientes pospago representa un 57% (6 millones) de la base total de abonados móviles de 10.5 millones. Es así, que el manejo de la base de clientes pospago que utilizan este servicio supone un gran impacto posterior en los ingresos y margen de Entel [7].

En esa línea, en lo que respecta a la tasa de abandono (fuga) del segmento de pospago, a diciembre del 2023 esta alcanzó un nivel de 1.86%, acumulando 2 trimestres debajo de su peak del 2023 de 2.06% [8], esto debido a un aumento de precios ocurrido en el primer trimestre (producto del panorama económico descrito anteriormente). No obstante, esta tasa sigue siendo mayor en comparación al 4to trimestre del 2022 (1.63%), por lo que aún no se alcanzan niveles deseados. En el Anexo E se pueden ver detalles de estos valores mencionados. En esa línea, para comprender mejor el fenómeno, es que Entel desde el año 2023 reporta mensualmente una encuesta de satisfacción en donde una muestra de los clientes fugados expresa su motivo de fuga.

1.2 Aspectos técnicos y de mejoras en redes

La señal que ofrecen las redes de telecomunicaciones de Entel se da mediante sus antenas, las cuales se encuentran ubicadas en lo que se llama un POP (Point of Presence o torre). Cada antena está conformada por celdas, las cuales son la unidad celular de estas y mediante la cual un dispositivo se conecta cuando realiza alguna operación en la red (llamada de voz, navegación por internet, etc.). En ese sentido, cuando un cliente se conecta a una celda es equivalente a decir que se conecta a la antena o al POP a la cual pertenece. Cada celda opera bajo una tecnología particular y posee distintos atributos relacionados principalmente a la velocidad de carga y descarga de datos, su utilización, su demora de transferencia, etc. En el Anexo F se puede ver un esquema que resume la estructura POP-celda-antena.

Algunos atributos considerados críticos para la experiencia o satisfacción de los clientes (y por ende en su propensión a fugarse) navegando por la red en un área de cobertura determinada tienen que ver con la velocidad de descarga de datos, el retraso de estos (latencia) y la velocidad de carga de datos (videos, servicios de streaming, páginas web, etc.). Sin ir más lejos, Entel fue galardonado el año 2023 como el mejor servicio móvil del país [9] debido precisamente por su rendimiento de estos atributos. Por lo tanto, las mejoras realizadas en redes móviles van apuntadas principalmente a mejorar el rendimiento de estos atributos en las antenas, dado el tráfico que recibe cada una día a día y las exigencias de los usuarios al navegar en ella. Otras mejoras orientadas a mejorar la experiencia de los usuarios pueden ser añadir una nueva banda 4G (en zonas con baja cobertura), añadir tecnología 5G, mejorar la capacidad instalada de la antena, etc.

2. Problema y proyecto

2.1 Descripción del problema

Durante el año 2023, la inversión bruta de Entel en Chile alcanzó los 342 mil millones de pesos [10], en donde un 60% estuvo orientada a servicios móviles [3], específicamente, destinada a infraestructura de red con el objetivo tanto de desplegar la red 5G en todo el país como de seguir fortaleciendo la red 4G, de forma de mantener el liderazgo durante la transición tecnológica de 4G a 5G [11]. Este escenario de inversión se espera se mantenga en 2024, teniendo en cuenta uno de los pilares de la estrategia corporativa de Entel relacionada a ser líderes en las tecnologías y telecomunicaciones móviles, ofreciendo a sus clientes una experiencia distintiva. Es por esto, considerando el alto porcentaje invertido en redes, que es muy importante para la reputación y sostenibilidad de la empresa dirigir de forma eficiente estos esfuerzos, representando cada año una oportunidad considerable para sacar los mejores resultados con ello, teniendo en cuenta además el impacto visto en la sección anterior que los clientes abonados pospago tienen en los ingresos de Entel.

En ese sentido, a pesar del buen estatus en participación de mercado, la empresa informa que cada mes se fugan aproximadamente 50 mil clientes pospago desde Entel hacia otras compañías a través de la portabilidad móvil. A algunos clientes les motiva tener un precio más bajo, o un servicio al cliente mejor, o descuentos en otras compañías que no se tienen en Entel. Esta fuga representa un obstáculo en la intención de seguir repuntando la tasa de abandono luego de su peak alcanzado en 2023. Dentro de los clientes que se portaron con su número a otra compañía (desde ahora hacer port out), desde la empresa se desconocen los motivos de sus fugas, es sólo a partir de una encuesta de satisfacción enviada a una muestra de fugados (aproximadamente un 1%) y de la experiencia, que se sabe que cerca de un 30% de ellos lo hacen por motivos de una mala experiencia de red, en otras palabras, por fallas técnicas de antenas (conectividad lenta o intermitente, baja señal, o cortes en sus llamadas, principalmente), lo cual representa un notorio quiebre con la estrategia y la imagen corporativa de Entel relacionada a ser líderes en la experiencia del cliente. En adición, estos clientes con mala satisfacción de red (a partir de la misma encuesta) se sabe que, en su mayoría, se portan a WOM o a Movistar (ver Anexo G), los cuales son los 2 competidores con más fuerza reciente de Entel acorde a la Ilustración 1.

A partir de esto, es importante para Entel saber cuáles son los clientes que se van por razones de mala experiencia de red (o fallas técnicas de antenas o de red) y cuáles son los POP en donde se concentran, para así con ello poder complementar la visión que se tiene sobre cómo dirigir de forma más eficiente el alto porcentaje de inversión en la red a la hora de priorizar los POP a mejorar, la cual actualmente no está considerando el cómo afectan estas intervenciones a nivel de clientes fugados. Todo esto con el objetivo de disminuir la fuga de clientes y seguir ofreciendo una experiencia de calidad a los clientes alineada con su estrategia y su imagen, considerando además que, en general, el costo de adquirir un nuevo cliente es entre cinco y seis veces más alto que retener uno ya existente [12].

En conclusión, una continuación de esta fuga pospago por mala experiencia sólo ocasionaría un mayor golpe a los resultados financieros de Entel, su reputación e imagen.

2.2 Descripción y justificación del proyecto

El proyecto que se realiza tiene por objetivo desarrollar un producto analítico que sea capaz de separar la base de clientes que hacen port out (desde ahora también hacer churn) en clientes que lo realizan por razones de mala señal o cobertura (por red) y no (otros o comerciales). El output de este proyecto permitirá identificar a los clientes fugados que tienen una mayor propensión a haberse fugado por razones de red, así como tener una mejor visión de los POP en donde se concentran y apoyar el trabajo del equipo SmartCapex de Entel, el cual tiene por objetivo dirigir y priorizar las inversiones de la empresa.

Considerando lo expuesto en la sección anterior, el proyecto se realiza con el motivo de complementar las visiones que se tienen actualmente desde el área de SmartCapex para priorizar las inversiones. En ese sentido, esta visión implica poder visualizar cuáles son los clientes que se fugan por mala satisfacción de red y a qué POP se les puede atribuir el mayor porcentaje de esta propensión. Además, desde Entel se ha estimado en la fase de ideación del caso de negocio del proyecto, que la mejora de un POP (explicado con más detalle en la sección 1.2) con esta visión podría disminuir el churn en 30 clientes mensuales, llegando así a niveles de fuga deseados (similares a los del cierre de 2022).

Para generar el ouput esperado del proyecto, se realiza un modelo de machine de learning de clasificación binaria que sea capaz de predecir (o atribuir) si un cliente pospago de Entel que hizo port out lo hizo por razones de fallas técnicas en redes o no, esto basándose en variables de distinta naturaleza (comerciales, kpis de experiencia, tráfico de datos, características de las celdas, etc.) que sean capaces de relatar el comportamiento y la experiencia previa del cliente usando las redes de Entel hasta cuando se portó. Además, con lo que el modelo prediga para cada cliente, se construyen las tasas de propensión de fuga por red promedio por cada POP, pudiendo observar así cómo se concentran. En esa línea, el proyecto necesita de un procedimiento que sea capaz de generar una asociación, posterior a la construcción y evaluación del modelo, entre cada cliente predicho y los POP.

La decisión de generar una solución analítica basada en datos (en este caso un modelo de clasificación binaria) viene de la mano con seguir el camino de la aspiración de Entel declarada de ser una empresa Data Driven, es decir, una empresa impulsada por datos y que crea valor a partir de la información disponible [13], aprovechando además las nuevas posibilidades que abre el uso de la inteligencia artificial en ello.

El proyecto es capaz de aportar y resolver de manera parcial el problema planteado debido a que, como se mencionó antes, actualmente no se está considerando una visión analítica en datos de cómo afectan las intervenciones de mejora a nivel de clientes fugados por calidad de red, por el contrario, algunas de las visiones (o criterios) tienen relación a, por ejemplo, análisis de fallas, consideración de degradación diaria de las antenas/celdas en la experiencia de los clientes, juicio de expertos, etc. Estas visiones son excluyentes entre sí, principalmente por el output que cada uno de estas tiene. Por lo tanto, al desarrollar el modelo y sus outputs, se abre la posibilidad de dar más completitud, confianza y respaldo analítico a la toma de decisiones de inversión dirigidas a la satisfacción de los clientes, aumentando así la eficacia operativa de la empresa y disminuyendo la fuga a niveles deseados (como se mencionó en el primer párrafo).

3. Objetivos y alcances

3.1 Objetivo general

Diseñar y construir un modelo predictivo que atribuya la fuga de clientes a razones asociadas a fallas técnicas de antenas y la asigne a los POP correspondientes.

3.2 Objetivos específicos

Para llevar a cabo el cumplimiento del objetivo general se plantean los siguientes objetivos específicos:

- Levantar la situación actual del proyecto y los insumos disponibles, tales como objetivo, tablas, variables, datos históricos de atribución, herramientas, etc.

- Construir y evaluar con datos fuera muestra un modelo de predicción de si el motivo de fuga de un cliente se debe o no a fallas técnicas de redes (o mala experiencia).

- Visualizar la propensión de fuga por red de cada cliente fugado, así como la concentración de tasas promedio de propensión de fuga por este motivo en cada POP y su cantidad de clientes asociados.

- Evaluar la consistencia y potencial estabilidad en el tiempo del modelo que permita dar una mayor confianza de adopción, junto con documentar la lógica del procedimiento, el desarrollo y los principales resultados del trabajo.

3.3 Alcances

Los datos utilizados tanto en el entrenamiento y la predicción corresponden a la base sólo de clientes (pospago) que realizaron port out, acorde al objetivo del proyecto de predecir sobre clientes ya fugados, en comparación a utilizar toda la base de clientes pospago. El tamaño de esta base representa entre el 1 y 2% del total de clientes abonados, lo cual va acorde con lo mencionado en la sección 1.1.3. Por otro lado, los datos para construir el modelo corresponden al 1% de la base total de churn mes a mes, esto porque considera una muestra a los cuales se les envió una encuesta que registra el motivo de su fuga, estos datos están desde enero a diciembre de 2023, esto principalmente porque la encuesta se comenzó a reportar desde ese año.

En términos de alcance de datos, el proyecto se centra en variables relacionadas más bien a la experiencia de los clientes en la red, como variables relacionadas a tráfico, tiempo de tráfico, atributos de celdas más traficadas, retraso de la señal, etc. esto se realiza con el fin de mantener la consistencia entre lo que se quiere predecir y las características a considerar, pudiendo traducirlas en accionables de mejoras. No obstante, el proyecto también recibe variables de tipo comercial interpretadas como relevantes para el objetivo de separar la base de clientes de port out, las cuales fueron trabajadas en etapas previas.

Ahora bien, existen datos que no están al alcance del proyecto, debido principalmente a su tamaño y costo en adquirirlos, estos datos tienen relación a fracciones de tiempo de tráfico de los clientes en las celdas que utilizó durante un período de tiempo determinado y a la duración de sus sesiones en cada una de estas. Esta limitación se maneja procesando los datos usados de manera que puedan generalizar a los no usados. Dentro de lo que compone un proyecto de data science en Entel (ver Anexo H), en este trabajo se aborda únicamente la etapa de MVP o Producto Mínimo Viable, esto principalmente debido a que las restantes están fuera del periodo de tiempo dedicado al proyecto. Esta etapa implica la validación con métrica fuera de muestra del modelo de predicción, dejando su desarrollo en estándares de la empresa con ejecuciones de entrenamiento y predicción.

En ese sentido, el proyecto parte con una ideación del caso de negocio y una PoC (etapas 1 y 2 en Anexo H) ya realizadas y validadas. Así mismo, el proyecto no involucrará una implementación en vivo ni una validación de impactos reales en la fuga e inversión (etapa Piloto en Anexo H). Por último, tampoco involucrará una automatización (etapa del mismo nombre en Anexo H) del modelo.

Finalmente, el proyecto llega para completar un vacío que se tiene en la visión global de la situación. Por lo tanto, si bien el proyecto apoya en las decisiones, es altamente probable que estas sean tomadas en el futuro a partir de la combinación de otras visiones desarrolladas por Entel (detalladas en la sección 2.2) y/o de otra iteración del trabajo.

Para mitigar estas limitaciones de etapas y adopción, se realizan pruebas de estabilidad del modelo que sean capaces de mostrar el cumplimiento de esta etapa, entregando una confianza tanto a las siguientes etapas como a la contraparte de adoptarlo.

4. Marco conceptual

A modo de introducir la metodología, se explican sus conceptos técnicos más relevantes.

4.1 Modelamiento predictivo con machine learning

En primer lugar, el machine learning es la ciencia de desarrollo de algoritmos y modelos estadísticos que utilizan los sistemas de computación con el fin de llevar a cabo tareas sin instrucciones explícitas, en vez de basarse en patrones e inferencias [14]. Por otro lado, un modelo predictivo es un modelo matemático o estadístico utilizado en inteligencia artificial y machine learning para predecir el valor de una variable de interés basándose en datos históricos y patrones observados. Normalmente el evento a predecir es futuro, sin embargo, el modelado predictivo se puede aplicar a eventos desconocidos sin importar el tiempo. Los modelos predictivos se utilizan en una variedad de aplicaciones, como el análisis de riesgos, la detección de fraudes, la predicción de la demanda, entre otros. [15].

Existen dos tipos de modelos predictivos: modelos de clasificación y de regresión [16].

- Los modelos de clasificación permiten predecir la pertenencia a una clase (ya sea binaria, de 2 clases, o múltiple, de más de 2). Por ejemplo, si se trata de clasificar entre clientes quiénes son más propensos al abandono, los resultados del modelo son binarios, o un sí o un no (en forma de 0 y 1) con su grado de probabilidad.

- Los modelos de regresión en cambio permiten predecir un valor. Por ejemplo, cuál es el beneficio estimado que se obtiene de un determinado cliente (o segmento) en los próximos meses o ayudan a estimar un forecast de ventas.

Para efectos del objetivo del proyecto, la técnica de modelamiento predictivo apropiada termina siendo una de clasificación binaria (el cliente se fue o no por motivos de red).

Ahora bien, referente a la literatura disponible, no se encontraron investigaciones que trataran con un fenómeno de fuga dada como en este proyecto. No obstante, en otra investigación, se probaron cuatro algoritmos distintos para predecir el churn de una empresa de telecomunicaciones: Árboles de Decisión, Random Forest, Gradient Boosted Machine Tree “GBM” y Extreme Gradient Boosting “XGBOOST” [17]. De aquí se obtuvo que aquel que presenta mejor desempeño es el último mencionado. Para entender estos algoritmos, es necesario definir lo que es un modelo de árbol de decisión, el cual es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente [18].

En ese sentido, un Random Forest es lo que se llama un método de conjunto (o ensemble method, en inglés), es decir que combina resultados para obtener uno final [19], en este caso, los resultados individuales son distintos árboles de decisión formados cada uno por

un subconjunto aleatorio de características. Por otro lado, tanto lightGBM y XGBoost son modelos que utilizan la técnica de aumento de gradientes, esto para entrenar a través de una secuencia de modelos que aprenden de su predecesor [20], la diferencia entre ambos radica principalmente en el método intrínseco que cada uno utiliza para ello.

4.2 Entrenamiento de modelos, sobreajuste y validación cruzada

En primer lugar, hay que saber que para entrenar un modelo y evaluar su capacidad de generalización usualmente se utilizan 3 sets de una data general (también llamado tablón maestro): Entrenamiento, validación y prueba. El set de entrenamiento permite obtener los parámetros (coeficientes numéricos internos del modelo y que se encuentran durante el proceso de entrenamiento) del modelo, y el de validación permite ajustar los hiper-parámetros (también coeficientes numéricos del modelo, pero que se definen buscando las mejores predicciones posibles), mientras que el set de prueba (o test) permite medir la capacidad de generalización del modelo [21]. Luego, el entrenamiento de un modelo predictivo involucra el uso del conjunto de datos de entrenamiento para "enseñar" al modelo a hacer predicciones precisas en el set de prueba [21]. Es importante mencionar, que el entrenamiento en cantidades de datos limitadas puede detener al modelo de aprendizaje automático de alcanzar su máximo potencial y aumentar el riesgo de proporcionar predicciones erróneas [22], luego, si el tamaño de los datos de entrenamiento puede ser aumentado, es recomendado hacerlo.

Ahora bien, es necesario indagar en una de las principales dificultades de un modelo de predicción, el llamado sobreajuste (u overfitting). El sobreajuste ocurre cuando un modelo estadístico se ajusta exactamente a sus datos de entrenamiento. Cuando esto sucede, el algoritmo no puede ser preciso contra datos invisibles (set de test) [23]. Si el modelo se entrena durante demasiado tiempo con datos de muestra o si es demasiado complejo, puede aprender el "ruido" o información irrelevante dentro del conjunto de datos, lo cual generalmente causa el sobreajuste [23]. En el contexto del proyecto, el sobreajuste puede ocurrir si se entrena demasiado el modelo con los datos de clientes de un solo periodo.

Una de las técnicas de entrenamiento más utilizadas para combatir el sobreajuste es la llamada k-fold cross-validation (o validación cruzada de k iteraciones), la cual tiene la ventaja de utilizar todo el set de entrenamiento, en lugar de una parte, como lo hacen otras técnicas [24]. Esta técnica consiste en particionar en k conjuntos de igual tamaño el set de entrenamiento y validación, luego, se realizan 3 pasos sucesivos: Se toma una de las *k* particiones y se mantiene “oculta” al modelo. Se toman las *k-1* particiones restantes y con ellas se entrena el modelo. Una vez entrenado, se almacena su desempeño para con el set de entrenamiento (*k-1* particiones) y con el set oculto. Luego, terminadas las iteraciones se tienen *k* medidas de desempeño para los sets de entrenamiento y validación usados en cada iteración. Así que el desempeño final del modelo será simplemente el promedio de los desempeños anteriores [25]. En el Anexo I se puede ver un esquema general de esta técnica.

Existen otras técnicas para combatir el sobreajuste, una de ellas es pausar el entrenamiento antes de que el modelo comience a aprender el ruido dentro del modelo, también conocido como "early stopping" o parada anticipada, o bien, se puede reducir la complejidad en el modelo eliminando entradas menos relevantes [23]. En el Anexo J se puede ver un esquema gráfico de cómo funciona la técnica “early stopping”.

4.3 Desbalance de clases

Otro problema en los modelos de clasificación es el desbalance de datos, esto ocurre cuando una clase o categoría es sesgada (hacia la clase positiva o negativa) [26]. El problema del desbalance es que la clase mayoritaria abruma a los modelos u algoritmos y desvían su rendimiento [27]. Dentro de los modelos mencionados anteriormente existen hiperparámetros que pueden manejar este desequilibrio. Tanto en Árbol de Decisión como en Random Forest existe el hiperparámetro class_weight, el cual se encarga de manejar los pesos de las clases [28]. Por otro lado, en los modelos GBM y XGBoost existe el hiperparámetro scale_pos_weight, el cual maneja el desequilibrio de clases al asignar un peso mayor a la clase minoritaria en comparación con la clase mayoritaria [29].

4.4 Métricas, visualización de resultados y optimización de hiperparámetros

Finalmente, los resultados de los modelos se evaluarán según distintas métricas, donde algunas de las más usadas en la disciplina de la data science se mencionan a continuación:

- Precisión: Mide la fracción de clasificaciones correctamente positivas con respecto al total de clasificaciones positivas del modelo [30].
- Especificidad: Mide la fracción de casos clasificados correctamente como negativos (con respecto al total de negativos reales) [30].
- Sensibilidad (o Recall): Mide la fracción de casos clasificados correctamente como positivos (con respecto al total de positivos reales) [30].
- AUC (Area Under Curve): Mide el área bajo la curva ROC (explicada posteriormente) y permite comparar un modelo con otro en todos los threshold, el cual corresponde al valor de probabilidad mediante el cual, si una observación lo supera, es clasificado como positivo o 1, y si no, 0. Esta métrica toma valores desde el 0 al 1, y modelos con un AUC mayor son considerados mejores clasificadores [30]. Se pueden desprender AUC de los distintos sets, en ese sentido, mientras mayor es el AUC de entrenamiento al de validación, mayor es el sobreajuste.

A partir de estas métricas, se utilizan las siguientes visualizaciones, las cuales tienen la

ventaja de permitir comparar entre diferentes modelos de propensión de una manera más directa mediante sus valores de probabilidad y sin necesitar un threshold [31]:

- Gráfico de valores shap: Muestra la contribución de cada variable a la predicción final del modelo. Las variables con valores shap positivos tienen un impacto positivo en la predicción, mientras que los que tienen valores negativos tienen un impacto negativo. La magnitud es una medida de la fuerza del efecto [32]. El eje X representa el valor shap, mientras que el eje Y representa si la observación tiene un alto o bajo valor, comparado con restantes de dicha variable [33]. El resultado de esta técnica permite estudiar las tendencias en el efecto de una variable, las interacciones y la influencia general que tienen sobre el output de un modelo. [30]

- Curva ROC (Receiver Operating Characteristic): Técnica de visualización que exhibe el trade off entre recall y especificidad de un modelo de clasificación al variar el umbral de clasificación. El eje Y corresponde a la sensibilidad, y su eje X corresponde a la resta de 1–especificidad [30]. Su uso deriva de la métrica AUC.

- Curva lift: Es una medida o la efectividad calculada como el ratio entre los resultados obtenidos con o sin un modelo. El eje X muestra el porcentaje de la población y se ordena de la probabilidad más alta a la más baja. El eje Y muestra cuánto es mejor su modelo que el modelo aleatorio [34].

- Curva precisión-recall: Resultado de dibujar la gráfica entre el precision y el recall entre distintos umbrales de clasificación. Esta gráfica permite ver a partir de qué recall se tiene una degradación de la precisión y viceversa [30]. En ese sentido, a partir de la intersección de estas 2 curvas se obtiene el threshold óptimo.

Por último, cuando se habla de optimización de hiperparámetros, este se refiere al proceso de encontrar la configuración de hiperparámetros que produzca el mejor rendimiento de un modelo [35]. En particular, existe la técnica Optuna, la cual consiste en un marco de software que identifica valores óptimos de hiperparámetro a través del método de prueba y error para un rendimiento eficiente del modelo, esto de forma automatizada y a partir de un set de valores iniciales para cada uno [36]. Su ventaja principal es que combina varios algoritmos al momento de realizar la búsqueda y optimización [37].

4.5 Intervalos de confianza

Por último, es necesario entender lo que es un intervalo de confianza (o IC). El IC describe la variabilidad entre la medida obtenida en un estudio y la medida real (o valor real) de la población. Corresponde a un rango de valores, cuya distribución es normal (ver Anexo K para entender cómo se ve esta distribución) y en el cual se encuentra, con alta probabilidad, la medida real. Esta “alta probabilidad” se ha establecido por consenso en 95%. Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el valor real de un parámetro con 95% de certeza [38]. En el Anexo L se puede visualizar la fórmula para determinar un IC bajo una distribución normal.

5. Metodología

A partir de los objetivos planteados la metodología comprende las siguientes fases y actividades:

- Levantamiento de la situación actual del proyecto
  -Entendimiento de fases previas desarrolladas y la actual
  -Revisión de datos y variables disponibles
- Desarrollo del modelo y sus resultados
  -Construcción del modelo mediante CRISP-DM
  -Desarrollo de resultados
- Análisis de consistencia y estabilidad en el tiempo del modelo
- Documentación del trabajo

De la mano con estas actividades, los lenguajes a utilizar son los siguientes:

VSCode: Editor de código abierto, su uso se da porque contiene una amplia gama de extensiones (como lenguajes Python, R y SQL), permitiendo realizar tareas de distinto tipo. No obstante, una de sus desventajas es que para proyectos con un alto volumen de datos o un código de ejecución pesada tiende a funcionar con una velocidad más lenta [39]. Además, la disposición del desarrollo es mediante distintos tipos de carpetas, en donde la principal es la que se denomina carpeta Pipelines, la cual cada una contiene un Pipeline que consiste en un conjunto de scripts ordenados que juntos desarrollan una determinada fase del proyecto (predicción, entrenamiento, procesamiento, ingesta, etc.).

Python: Este lenguaje se utiliza como herramienta para explorar, desarrollar y validar el modelo, guardando los códigos y sus outputs en distintos scripts ordenados en carpetas dentro de VSCode. El motivo de su uso es principalmente por su sintaxis clara y legible, además de su amplia biblioteca estándar en proyectos con datos.

GIT (control de versión): Su uso se da a raíz de abrir la posibilidad de que distintos usuarios puedan trabajar en un mismo proyecto a través de un repositorio virtual (en este caso uno denominado Bitbucket), el cual sirve para ir dejando constancia de las versiones del proyecto que uno realiza localmente.

SQL: Se usa desde la aplicación Athena de AWS, la cual es un Data Lake en la nube (ambiente en el cual quienes tengan acceso pueden ver su contenido) y es en donde se almacenan todas las tablas de interés para los distintos proyectos. El uso de SQL es para realizar las consultas de las tablas útiles y utilizarlas para extraerlas desde el Data Lake.

Power Point (PPT): Su uso es para generar bitácoras de avance, así como documentar el procedimiento y las conclusiones del trabajo.

Antes de pasar a la descripción de cada fase, en el Anexo M se muestra una carta Gantt con los plazos estipulados para el desarrollo de cada una, junto con sus actividades.

5.1 Levantamiento de la situación actual del proyecto

Esta primera fase tiene por objetivo entender la etapa actual en la que se encuentra el proyecto, así como los insumos que se encuentran disponibles para el trabajo, esto con el fin de llevar a cabo el cumplimiento del 1er objetivo específico.

5.1.1 Entendimiento de fases previas desarrolladas y la actual

Para realizar esta actividad se sostienen conversaciones con el equipo en el cual se encuentra inmerso, además de una revisión de la documentación disponible del proyecto, la cual comprende principalmente presentaciones en formato PPT de las etapas previas desarrolladas y sus conclusiones, así como el objetivo del proyecto, su impacto económico y la etapa en la que se encuentra actualmente.

Los insumos disponibles para esta actividad son principalmente una presentación en formato PPT del trabajo realizado previo y una presentación inicial de la etapa actual respecto al objetivo e impacto del proyecto.

5.1.2 Revisión de datos y variables disponibles

En esta actividad se revisan las tablas disponibles para el desarrollo del proyecto, considerando la información que proporcionan y a partir de dónde se consideran (desde una fase previa, recomendación del negocio o del área, etc.).

Para realizar esto, se llevan a cabo principalmente conversaciones con el equipo sobre el propósito de cada tabla, además de ciertos análisis exploratorios corroborando la información proporcionada como contexto del proyecto. Respecto al desarrollo de esta actividad, se utiliza el lenguaje Python en VSCode y SQL para la ingesta. Por último, en el Data Lake se cuenta con datos disponibles de las encuestas de satisfacción del año 2023 y las variables utilizadas en etapas anteriores del proyecto con su respectiva tabla.

5.2 Desarrollo del modelo y sus resultados

El objetivo de esta fase es generar el modelo de atribución (2do objetivo específico) para madurar el desarrollo del proyecto. Lo fundamental de esta etapa es tener una ejecución de modelamiento y predicción, validando métricas de rendimiento con datos fuera de muestra. En específico, las actividades de esta fase son las que se detallan a continuación.

5.2.1 Construcción del modelo mediante CRISP-DM

Comprende todo el proceso desde la construcción del tablón maestro a utilizar, su

procesamiento, el modelamiento y la validación (o evaluación) con datos fuera de muestra. La metodología a utilizar es CRISP-DM (Cross Industry Standard Process for Data Mining), utilizada comúnmente en proyectos de datos que permite volver al punto inicial a medida que se va avanzando. Esta metodología fue seleccionada para el presente trabajo debido a que ha mostrado un buen desempeño resolviendo proyectos de ciencia de datos independiente del sector industrial y tecnología del que se hace uso, donde provee tanto la estructura como la flexibilidad necesaria para ejecutar distintos proyectos de esta índole [40]. Esta metodología consta de 6 etapas: Entendimiento del negocio, entendimiento de la data, preparación de la data, modelamiento, evaluación y desarrollo.

Para los alcances de este trabajo (sección 3.3), la metodología sólo llegará hasta la etapa de evaluación, esto debido a que la etapa de desarrollo corresponde a todo lo relacionado a la implementación de la solución (etapas Piloto y Automatización de las fases de un proyecto en Entel en el Anexo H). Luego, las etapas de CRISP-DM de esta actividad, junto con la descripción de su realización en el contexto del trabajo, son las siguientes:

Entendimiento del negocio

En esta etapa, es definido el objetivo del proyecto y las necesidades de la empresa o proyecto en análisis [41]. En el proyecto en esta etapa se recopila información sobre el su objetivo basado en el caso de negocio, así como las actividades a realizar, esto mediante conversaciones con el equipo, investigaciones sobre terminologías y comprensión de los recursos disponibles (herramientas, datos, realizaciones previas, etc.). Esta actividad se deriva de la fase anterior y es crucial para el desarrollo del modelo en el sentido de que permite ser consistente con lo que esperado y con lo ya desarrollado.

Los principales insumos utilizados y combinados en esta parte son las presentaciones derivadas de la fase anterior, una planilla Excel del caso de negocio desarrollado y un documento Miró de los principales términos y atributos en redes.

Entendimiento de la data

Esta fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema [42]. En el proyecto en esta etapa se realiza toda la recolección de las tablas a utilizar desde el Data Lake disponible para ello y mediante consultas SQL. Luego, comprende la construcción del tablón maestro y su exploración (EDA o Análisis Exploratorio de la Data) utilizando lenguaje Python en VSCode. A partir de esta etapa se pueden tener variables disponibles para desarrollar el modelo.

Preparación de la data

Es la fase intermedia para seleccionar, limpiar y generar conjuntos de datos correctos, organizados y prepararlos para la fase de modelado [42]. En el proyecto en esta etapa se limpian, formatean y seleccionan las variables a utilizar. Esto comprende tratamiento de valores nulos, de outliers, variables categóricas, posibles ruidos, creación de variables a partir del EDA, etc. Esta etapa es crucial dado que la calidad y relevancia de las variables usadas para entrenar el modelo son fundamentales para evitar el sobreajuste y acelerar su

rendimiento computacional (ejecución), facilitando su mantención. Si el conjunto de variables contiene datos ruidosos, irrelevantes o incluso redundantes, el modelo puede aprender a ajustarse a esas variables erróneas en lugar de capturar los patrones genuinos [43], además, resulta en una oportunidad para reducir el número de variables y acelerar la ejecución. El desarrollo de esta etapa se realiza en el lenguaje Python en VSCode.

Modelamiento

En esta fase se lleva a cabo la creación de modelos a partir de los datos suministrados desde la fase anterior [42]. En el proyecto en esta etapa se aplican los algoritmos predictivos de machine learning, tales como Random Forest, lightGBM y XGBoost (siguiendo lo expuesto en [17]), con el objetivo de atribuir la fuga. En esa línea, en esta etapa se realiza el entrenamiento del modelo, lo cual comprende separar la base en sets de entrenamiento, validación y test, elegir un modelo, construirlo, iterarlo, optimizarlo y reentrenarlo, todo este proceso se realiza dentro del lenguaje Python en VSCode. El mayor cuidado y lo crucial de esta fase es el de evitar lo más posible el sobreajuste.

Evaluación

Para esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema [42]. En el proyecto, en esta etapa se evalúa el modelo mediante una validación de sus predicciones en el set de test a partir de la métrica AUC, esto aprovechando la robustez que posee y la posibilidad que ofrece de comparar entre modelos (sección 4.4). En ese sentido, se procura que los AUC de entrenamiento, validación y test se asemejen lo más posible, con el fin de minimizar el sobreajuste. En esta etapa se valida que el modelo no pierda precisión en datos que no ha visto durante el entrenamiento. El desarrollo de esta etapa se realiza en el leguaje Python en VSCode.

5.2.2 Desarrollo de resultados

En esta actividad, se construyen los entregables principales que se espera que el modelo genere (3er objetivo específico), en este caso, la visualización de la propensión de cada cliente fugado a hacerlo por motivos de red y las tasas promedio de propensión por POP, junto con su cantidad de clientes asociados a partir de la asignación de la atribución.

Para esta demostración, se utilizan inicialmente los meses utilizados en la etapa de evaluación como predicciones iniciales de la base total de fuga. En ese sentido, se construye una tabla a nivel mensual que muestre 3 atributos: El cliente, su probabilidad de atribución de su fuga por red entregada por el modelo y la clase predicha (aplicando el threshold óptimo derivado de la curva precisión-recall).

Por otro lado, se construye una data indexada por POP, la cual posea la comuna en donde está, su cantidad de clientes fugados asociados, y su probabilidad promedio de propensión de fuga por red. Para ello, se utiliza un criterio de asignación de los clientes a los POP.

Para esta actividad se cuenta con datos disponibles en el Data Lake referentes a los clientes

totales que se fugaron mediante portabilidad mes a mes, además de variables asociadas a su experiencia y utilizadas en el modelo.

5.3 Análisis de consistencia y estabilidad en el tiempo del modelo

Esta fase comprende una evaluación de la robustez en el tiempo del modelo en relación a la estabilidad y consistencia de sus resultados generados (4to objetivo específico), esto con el fin de generar una mayor confianza de adopción en la contraparte del modelo como producto a implementar. Estos análisis son en base a recomendaciones del equipo, de la contraparte y en apoyar un potencial desarrollo (etapa 6 de CRISP-DM) del modelo.

En primer lugar, se realizan predicciones en toda la base de clientes de port out en un horizonte de más de 1 mes y se compara el promedio de atribución de fuga por motivos de red con lo esperado en base a un intervalo de confianza al 95%. Luego, se compara la cantidad de POP críticos coincidentes entre estos meses en base al cruce entre su tasa de propensión promedio de fuga por red y la cantidad de clientes asociados. Por último, se revisa el comportamiento entre estos meses de atributos críticos en la experiencia del usuario y su relación con las tasas de propensión, revisando su consistencia.

Para esta fase se cuenta con datos relacionados a los clientes fugados en cada mes considerado, así como los atributos de interés para las celdas. Esta fase se realiza en el lenguaje Python en VSCode para desarrollar los scripts de código y en el lenguaje SQL en Athena para ingestar los datos, todo esto dentro del proceso 'postproessing' (o postprocesamiento) definido en el Pipeline de desarrollo como una carpeta de scripts.

5.4 Documentación del trabajo

Finalmente, esta fase comprende una documentación del procedimiento (con su razonamiento detrás), las principales conclusiones y el desarrollo mismo del proyecto, esto con el objetivo de hacer más entendible el trabajo para aquellos agentes que lo utilicen posteriormente (tales como la contraparte, equipo de automatización, etc.) y de complementar la sección 5.3 para el cumplimiento del objetivo específico 4.

Esta actividad comprende la documentación del trabajo y sus resultados en una bitácora del proyecto en formato PPT, en donde con fecha declarada se irá dejando constancia sobre los resultados/insights obtenidos junto con su procedimiento asociado.

En adición, como se mencionó anteriormente, se utiliza el repositorio Bitbucket para ir dejando las versiones del código respaldadas. Por último, se aprovecha la ventaja de VSCode en disponer el desarrollo del modelo desde la ingesta de datos hasta evaluación de la estabilidad en carpetas Pipeline. Estas 3 aristas de documentación se respaldan en los estándares utilizados por la empresa en proyectos con datos.

6. Desarrollo y Resultados

6.1 Levantamiento de la situación actual del proyecto

6.1.1 Entendimiento de fases previas desarrolladas y la actual

A partir de lo recopilado, el proyecto ha llevado a cabo ya 2 etapas: Ideación del caso de negocio y prueba de concepto (o PoC). De forma cronológica con el trabajo, se analizó primeramente la PoC del proyecto. Esta consistió en validar si técnicamente era factible el objetivo del proyecto. Este desarrollo concluyó con lo que se puede ver en las figuras (a) y (b) de la Ilustración 2, las cuales muestran la métrica AUC de entrenamiento y validación (dentro de muestra), respectivamente, del modelo de clasificación generado. Además, en el Anexo N se puede ver el gráfico de valores shap obtenido.

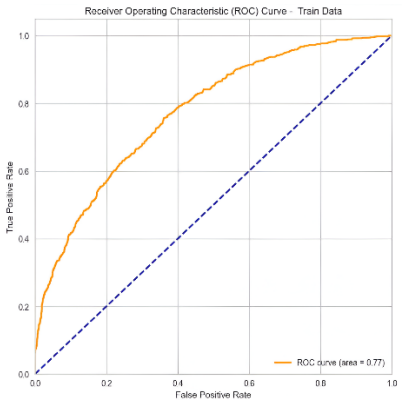


Figura (a): AUC de entrenamiento (0.77)

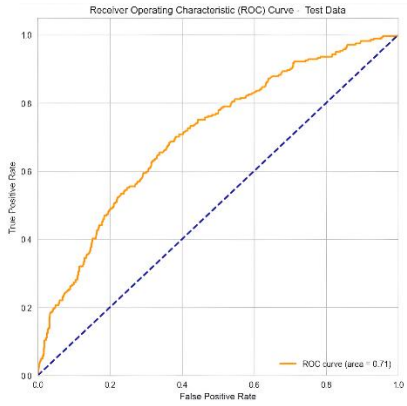


Figura (b): AUC de validación (0.71)

Ilustración 2: Curvas ROC en entrenamiento y validación

Cabe mencionar, que el AUC de validación del modelo alcanzó valores entre 0.68 y 0.72 a partir de las distintas iteraciones de entrenamiento, lo cual termina siendo consistente con investigaciones relacionadas a propensión de fuga en otras áreas de Entel [44] y resultó aceptable para la contraparte al momento de presentarlo, por lo que se utiliza este marco de valores para comparar la métrica AUC del modelo desarrollado en este informe.

El proyecto actualmente se encuentra en la etapa de MVP, el cual consiste en generar e iterar el modelo de la PoC en códigos limpios de entrenamiento y predicción y acordes a estándares de Entel, evaluando en fuera de muestra y realizando las pruebas de confianza pertinentes antes de implementar. Además, se da cuenta de la necesidad de variar el rango de meses considerados en la PoC, el cual termine siendo considerar ahora a los clientes fugados de marzo a diciembre del 2023. Esta necesidad surge de actualizar los datos considerados, dado que en la PoC se consideraron datos de fuga hasta octubre de 2023

(por el tiempo en el que se realizó). En adición, no se consideran enero y febrero, por ser meses sujetos a anomalías en el comportamiento (comúnmente de vacaciones), el impacto de esto se revisará más adelante. En el Anexo O se muestra una lámina resumen de esta actividad, proveniente de la bitácora de la etapa actual del proyecto en formato PPT.

Las principales conclusiones de esta actividad son las siguientes:

- La solución planteada al problema y el fenómeno a estudiar es factible técnicamente de abordar, luego, el trabajo que se desarrolla no tiene riesgos a priori de no funcionar, dado que itera sobre lo ya realizado.

- Se rescatan variables relevantes de este análisis tal como se muestra en el Anexo N, junto con recomendaciones de agregar otras, a partir del feedback recibido.

- Se obtiene un marco de referencia de comparación del rendimiento del modelo a desarrollar, el cual en este caso es de un AUC de test de 0,70 +- 0,02.

- Se comprende la nueva configuración del marco de tiempo considerado en datos para la etapa actual, en base al objetivo de esta y del proyecto.

6.1.2 Revisión de datos y variables disponibles

En primer lugar, se cuenta con una tabla llamada port_out_survey, la cual representa una muestra de la población de clientes (aproximadamente un 1%, resultante en 550 clientes mensuales en promedio), que realizaron port out en el período de marzo a diciembre del 2023 y a los cuales se les envió una encuesta de satisfacción, en donde entre otras cosas declararon si el motivo de su fuga fue por red o no, considerando este atributo finalmente como el target del modelo (variable a predecir). En el Anexo P se puede ver un vistazo a esta tabla, en donde la columna 'motivo' hace referencia al target de cada cliente.

Luego, en la Ilustración 3, se aprecia la distribución del target en los meses de la encuesta.

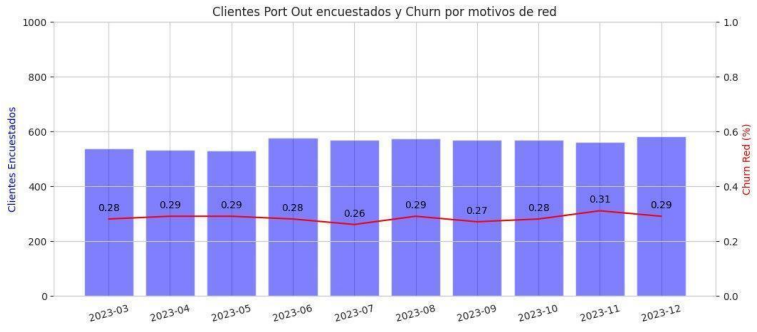


Ilustración 3: Distribución de la tasa de fugados por red en los períodos de encuesta

A partir de esto, se obtiene un promedio del 30% de clientes mensuales en esta muestra que se van por motivos de red, derivando en lo expuesto en la sección 2.1.

En segundo lugar, se cuenta con tablas y variables ya disponibles y consultadas para las etapas de entrenamiento y evaluación del modelo. A continuación, en la Tabla 1 se presenta cada dimensión de variables, junto las tablas consideradas. Además, en el Anexo Q se puede ver la descripción en detalle de las variables de cada una de estas:

Tabla 1: Dimensiones de variables a considerar y sus tablas incluidas

Dimensión	Tablas
Intensidad y calidad de la señal 3g y 4g recibida por los clientes	bt_ecno_rscp bt_rsrq_rsrp
Kpis celdas (como NUT y PRB)	redes_bt_calidad_3g_diario redes_bt_calidad_4g_diario
Actividad de tráfico en internet de los clientes	cdr_bt_datos_kpis_experiencia
Celdas favoritas y tasas de uso por tecnología de los clientes	cdr_bt_kpi_trafico_diario
Calidad de la navegación web	huawei_sdr_msisdn_ps
Antigüedad de los clientes	bi_fct_clnt_pspg_bgd
Datos geoespaciales de las celdas	bt_redes_celdas

Excepto la tabla huawei_sdr_msisdn_ps, el resto fueron consideradas a partir del trabajo realizado en la PoC, la cual las consideró todas ellas para construir el modelo. Respecto a la tabla mencionada, esta se consideró a partir de la recomendación de la contraparte del área, la cual sugirió que las variables que relatan la experiencia de navegación también son fundamentales para discriminar una mala experiencia de red. Variables relacionadas a aspectos más comerciales de los clientes, sus dispositivos, actividad de llamadas, edad, etc. no fueron consideradas ya sea por no ser relevantes desde la PoC o por no ser mencionadas desde el equipo y la contraparte. Por otro lado, a excepción de las tablas bt_redes_celdas, redes_bt_calidad_3g_diario y redes_bt_calidad_4g_diario, la cantidad de registros de las demás comprende la cantidad de clientes encuestados en cada mes de entrenamiento (del orden de 500 clientes). Esta diferencia es debido a que estas 3 tablas son a nivel de celdas, mientras que el resto se obtiene a nivel de clientes.

6.2 Desarrollo del modelo y sus resultados

6.2.1 Construcción del modelo mediante CRISP-DM

6.2.1.1 Entendimiento del negocio

En primer lugar, se comprenden los conceptos claves a considerar en una etapa de inducción, en donde a partir de sesiones de aprendizaje, investigaciones autónomas y sitios de inducción de Entel, se estudia lo que consiste una red de telecomunicaciones, así como se comprende el uso de Bitbucket (con GIT), AWS y VSCode en el desarrollo de un

proyecto. En lo que respecta a la ideación del caso de negocio, este consistió en explorar si el caso de uso era viable económicamente. En ese sentido, la revisión inicial que se realizó como posible impacto del proyecto es que la mejora de un POP mediante el criterio desarrollado podría disminuir el churn en 30 clientes mensuales, recuperando la inversión de dicha mejora en un plazo de 22 meses. Este valor se calcula bajo el supuesto de que la fuga anual promedio asociada a los POP sea igual al promedio nacional obtenido durante el 2023, el cual es de un 1,67% (asemejando la tasa de fuga al cierre del 2022 de un 1,63%. En el Anexo R se puede visualizar un vistazo a la planilla del caso de negocio.

Finalmente, mediante conversaciones con el equipo y la revisión de la presentación del Anexo O, se concluyen los siguientes puntos:

- La contraparte es el equipo de inversiones de SmartCapex, en donde el negocio es poder disminuir la fuga de clientes en base a mejores en la infraestructura de los POP, con criterio de ver cuáles son aquellos en donde se concentra esta mala satisfacción y tienen asociados un mayor número de clientes.

- Se cuenta con datos asociados a la experiencia en redes del cliente, pero no de tipo, por ejemplo, financieros. Además, para posibles análisis de consistencia, se pueden utilizar 2 atributos de celdas críticos recomendados por la contraparte: NUT (velocidad de descarga de datos) y PRB (utilización).

- El modelo desarrollado sirve y agrega valor encontrando POP críticos, en el sentido de identificar a los clientes propensos y asignándolos a los POP, además de poseer una robustez en el tiempo para eventualmente implementarlo en producción.

- El modelo necesita quedar documentado en desarrollo y procedimiento bajo estándares de la empresa (descritos en la sección 5.4), para ofrecer mayor claridad a la contraparte y que puedan ejecutar posteriormente.

- Su implementación va dirigida para el último trimestre de cada año (en particular, octubre y noviembre), apoyando el armado del plan de priorización de inversión para el año siguiente y posterior a la realización (o no) de más iteraciones del modelo hasta ese momento, derivadas, principalmente, de los feedbacks recibidos.

6.2.1.2 Entendimiento de la data

6.2.1.2.1 Horizonte de tiempo y creación de data

En primer lugar, se decide el horizonte de tiempo como historia para cada cliente. Esto se realiza considerando previamente las siguientes sugerencias del equipo y de la consultora Mckinsey, la cual está desarrollando en el área proyectos paralelos de experiencia en red.

- La convención de historia está en el rango de 30 a 90 días (1 a 3 meses) antes del churn. Esto por el tiempo de procesamiento en considerar un mayor horizonte y por la posible indisponibilidad de datos productivos de las tablas utilizadas.

- El aporte marginal de información al indexar de forma semanal o diaria en

comparación al hacerlo de forma mensual no es lo suficientemente superior como para merecer un gasto en procesar esa cantidad de datos, esto sumado a que no se sabe con certeza en qué momento exacto del mes a predecir se efectuó el churn, principalmente porque la encuesta se encuentra a nivel mensual, y su respuesta se efectúa posteriormente al momento de hacer churn.

Con estas consideraciones, se realiza la evaluación (de 1 a 3 meses) del aporte de información de las variables relacionadas a la tabla cdr_bt_datos_kpis_experiencia, esto a partir de recomendaciones del equipo dirigidas a que estas variables representan de forma más clara el comportamiento de los usuarios navegando por la red. A partir del ranking de correlaciones con el target que se puede ver en los Anexos S (a) y (b) por ambos lados, se puede concluir que las variables en general terminan siendo más alineadas con este en el mes anterior a la fuga (en cada grupo de variables de igual naturaleza), por lo que se decide considerar un mes de historia en las variables.

La creación de la data (o tablón maestro) se realiza considerando un merge (unión) entre las tablas descritas en la Tabla 1. Esto se realiza siguiendo el esquema de la Ilustración 4:

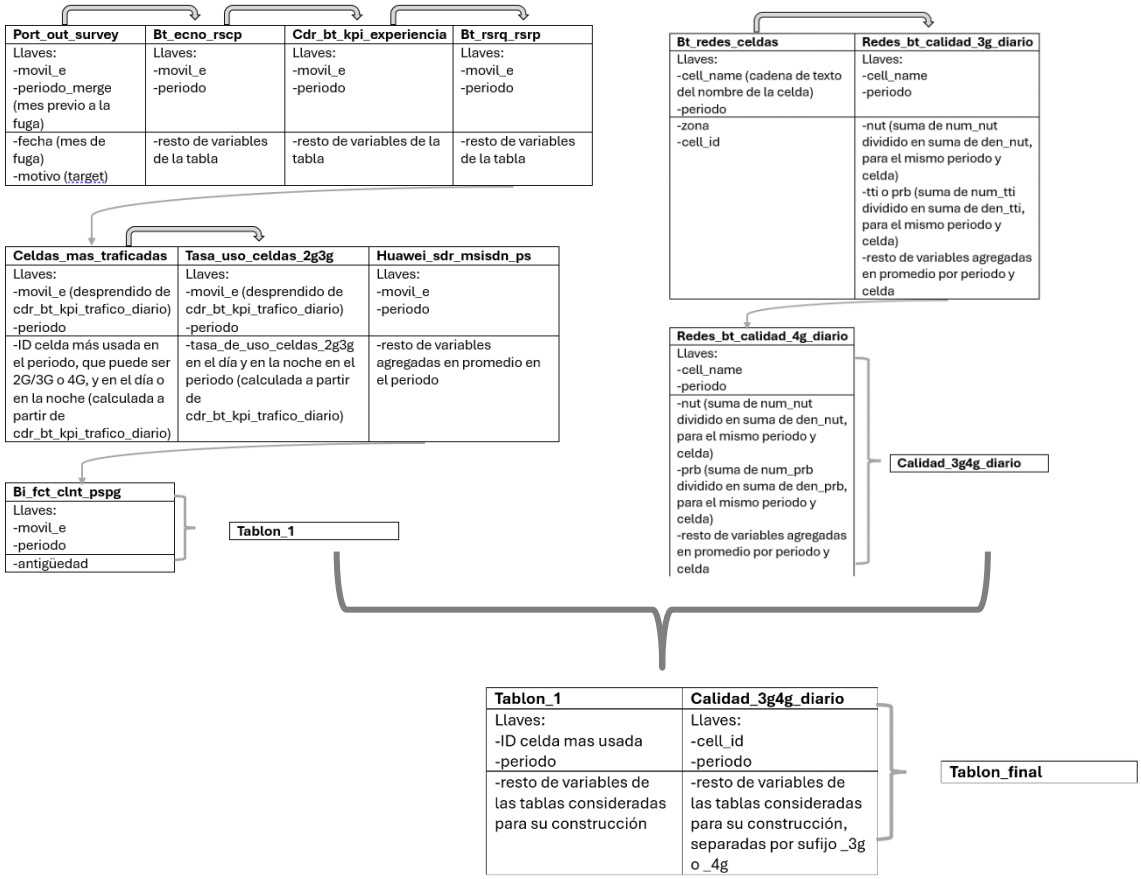


Ilustración 4: Esquema de armado del tablón maestro

Como consideraciones adicionales se menciona lo siguiente:

- Dado que 'periodo_merge', se refiere al mes previo del port out, las tablas consideran en cada uno de los merge periodos desde febrero a noviembre del 2023.

- Se consideran directamente las tablas bt_ecno_rscp. bt_rsrq_rsrp, cdr_bt_kpi_experiencia en el merge bajo las llaves de 'movil_e' y 'periodo_merge', esto porque ya poseen la información a nivel mensual.

- A partir de la tabla cdr_bt_kpi_trafico_diario se calculan 2 variables a nivel mensual; tasa de uso 2G/3G en el día (cantidad de celdas 2G/3G en el día sobre cantidad total de celdas traficadas en el día) y en la noche (análogo, pero con las de noche), con motivo de ser ambas variables relevantes según la PoC en el Anexo N.

Por último, se da cuenta que la disponibilidad de datos de la tabla de calidad de navegación es desde abril, luego, se consideran clientes que hicieron port out desde mayo, obteniendo el tablón maestro final. El tablón resultante consta de 4495 filas y 104 columnas, cuyo target es la columna 'motivo'. Un vistazo desde el código resultante del tablón se puede ver en el Anexo T.

6.2.1.2.2 EDA (Análisis Exploratorio de la Data)

El target presenta un desbalance de un 28.38% y 71.62% a favor de la clase 0 (o motivo = 'Otros'), por lo que se hace necesario hacerse cargo de esto más adelante.

La data consta en total de 4482 clientes únicos, por lo que existen clientes que tienen registros duplicados (un 0.2%). No obstante, esto no se considera un problema, dado que puede representar una situación en la que dicho cliente se fugó 2 veces dentro del mismo intervalo de tiempo en distintos meses, en ese caso, se considera la experiencia de cada cliente al mes previo como al resto. En conclusión, no es necesario tratar estos registros.

Un tema interesante de abordar es si existe alguna estacionalidad en la fuga. En esa línea, la opción de, por ejemplo, hacer otro modelo para considerar los meses descartados de fuga (enero y febrero), no se considera necesario dado que, como se dijo, el modelo no se implementa en periodos en donde se necesiten estos meses. Respecto a estacionalidades en relación a otros aspectos, no se consideraron análisis a nivel de edad, género u otro aspecto, principalmente por temas de tiempo para desarrollar el modelo, por lo que se deja para una segunda iteración fuera de los tiempos de esta memoria.

Aterrizando el EDA, se realiza un ranking de correlaciones entre las variables y el target, para tener un inicio con el análisis, en la Tabla 2 se puede ver un vistazo de esto:

Tabla 2: Top 10 de variables más y menos correlacionadas con una fuga por red.

Variable	Correlación con fuga por red
Tasa_noche_q_celdas_2g3g | Median_rsrp	0.219 | -0.167
Percent_umbral_rsrp | Scorecobertura4g	0.200 |-0,165
Tasa_dia_q_celdas_2g3g | p_tiempo_4g	0.192 | -0.162
Delay streaming 3g | P05_rsrp	0.166 | -0.160
Delay sitios web en 3g | Cell_avg_nut_dl_3g_noche	0.163 | -0.155
Cob4mala_rate_4g | P_tiempo_4g_activo	0.163 | -0,153
P_cambios_4g_3g | P_celdas_4g	0.157 | -0,145
Sm_ps_3g_chile_page_sr_delay_msel | P95_rsrp	0.155 | -0.131

Delay multimedia 3g | Cob4buena_rate_4g	0.153 | -0.127
Sm_ps_3g_page_response_delay_ms_kpi | Scorecobertura 3g	0.151 | -0,125

Acorde a la teoría, tiene sentido que las métricas relacionadas a uso de celdas 3G estén más correlacionadas con hacer churn por motivos de red, esto por lo explicado en secciones previas de que esta tecnología va más bien de salida. De este mismo modo, también hace sentido el hecho de que una mejor cobertura 4G, así como un mayor uso de celdas con esta tecnología, se relacione más con clientes que hacen churn por otro motivo.

Siguiendo, en los Anexos U (a), (b), (c) y (d) se muestran las proporciones entre cada clase del target para cada una de las categorías asociadas a la zona de las celdas más usada 3G día, 3G noche, 4G día y 4G noche, respectivamente. De estos gráficos se desprende que las zonas que tienen una mayor proporción de clientes que hicieron churn por motivos de red son las "No Urbanas" y "Pueblos", por lo que se podría separar.

6.2.1.3 Preparación de la data

6.2.1.3.1 Tratamiento de valores nulos

El tablón posee 96 variables con valores nulos, de los cuales 6 tienen un 67%, 18 un 40%, 6 un 34%, 3 un 12% y el resto posee un 10% o menor. Estos valores nulos se saben que se generan, principalmente, por la no disponibilidad de los datos, esto proveniente de errores en la ingesta de tablas a la nube o de la actividad misma de los clientes que no se puede representar mediante las otras variables. El tratamiento de estas variables se hizo mediante 2 criterios: Magnitud y varianza. Para aquellas variables que no superaban el 10% de sus registros como nulos, se optó por técnicas de imputación, en el caso de las numéricas por el promedio, mientras que en el caso de las categóricas por la más frecuente. Por otro lado, para aquellas variables que superaban dicho porcentaje se optó por eliminar aquellas cuyas varianzas se escapaban del promedio dentro de un mismo grupo de porcentajes, mientras que el resto se imputó. El motivo de utilizar una imputación por el promedio viene de la mano con que es el método más común utilizado con variables numéricas [45], además, el hecho de imputar las variables con varianza menor al promedio viene de considerar un criterio para separar aquellas variables con una "alta" y "baja" varianza, considerando que imputar variables con una alta varianza puede sesgar su análisis al no ser dicho promedio lo suficientemente representativo.

6.2.1.3.2 Transformaciones

A partir del EDA, se crean 4 variables binarias que indican si la zona de la celda más usada (ya sea 3G o 4G y en el bloque que corresponda) corresponde a "No Urbanos" o "Pueblos", o no. Como consecuencia, la data deja de contar con variables categóricas.

Además, a partir de su importancia en la PoC, se adicionaron 3 variables asociadas a un score ponderado del porcentaje de tiempo activo, porcentaje de tráfico y porcentaje de uso de celdas combinando las tecnologías 3G y 4G, esto dado que en la PoC estas variables fueron relevantes. La construcción de estas variables se realizó mediante la siguiente fórmula: Variable 4G x Score cobertura 4G + (1- Variable 4G) x Score cobertura 3G.

6.2.1.3.3 Tratamiento de valores outliers

En base a la experiencia en proyectos de data science, el tratamiento de valores outliers se realiza mediante el criterio del rango intercuartil. En esa línea, se opta por una imputación basada en los valores límites (o cotas), luego, cada valor que sobrepasa alguna de las cotas (ya sea superior o inferior) se imputa por la respectiva cota, y así para cada variable. La decisión de hacer esto es con el fin de no perder la interpretación de "magnitudes altas o bajas" de las variables. Este procedimiento resultó en 83 variables imputadas de 85.

6.2.1.3.4 Tratamientos adicionales

Siguiendo lo expuesto en la sección 5.2.1 del posible problema en considerar atributos ruidosos, irrelevantes o incluso redundantes, en primer lugar, se eliminan variables que estuviesen muy correlacionadas entre sí. En particular, de entre los 10 pares de variables más correlacionadas entre sí (correlaciones entre 0.9 y 1) se elimina en cada par la que menos correlacionada esté con el target. El motivo de este tratamiento es para acelerar el rendimiento computacional del modelo al reducir dimensionalidad (número de variables) [46]. En segundo lugar, el tratamiento de variables irrelevantes se realiza excluyendo las que tengan varianza próxima o igual a 0, esto porque variables con pocos cambios en los datos contienen en general poca información [47]. Luego, se opta por eliminar variables que tengan una desviación estándar (o SD) menor a 0.01, esto en base a mirar las variables con SD más baja dentro del último decil de estos. Al final de estas 3 etapas y previo al modelamiento, la data resultante consta de 4495 filas y 80 columnas (eliminando además periodo_merge). En el Anexo V se puede ver en detalles las variables eliminadas en cada tratamiento, junto con la tabla a la cual pertenecen (o si fue creación propia).

6.2.1.4 Modelamiento

En esta etapa se espera obtener un modelo que tenga una máxima precisión y mínimo sobreajuste, cuyos AUC de entrenamiento y validación no estén alejados por más de 0.1 (10 %) [48] y que este último al menos esté en el rango obtenido en la evaluación de la PoC (0.70 +- 0.02), si es que no lo supera.

6.2.1.4.1 Separación set de test

En primer lugar, se separa la data de test del entrenamiento mediante su aislamiento del tablón maestro, esta data corresponde a los registros del mes de diciembre, correspondiente a los clientes fugados en dicho mes y sus respectivas variables. El motivo de elegir este criterio es poder contar con la mayor cantidad de datos de entrenamiento (siguiendo lo expuesto en el marco conceptual en [22]), en la medida que se deje al menos 1 mes para la evaluación. Además, viene de poder considerar un rendimiento fuera de muestra que permite darle más robustez al modelo, algo que no se realizó en la PoC.

6.2.1.4.2 Modelo

A partir de los modelos revisados en el marco conceptual (en [17]) y la experiencia del equipo, se opta por probar 3 modelos principales: lightGBM, Random Forest y XGBoost.

Posteriormente, se decidió apuntar a este último dado que, aplicando iguales tratamientos de la data de entrenamiento y optimizando hiperparámetros en cada uno, obtuvo un mayor balance entre los valores AUC de los sets de entrenamiento, validación y test, respecto a los otros, lo cual termina siendo consistente con la investigación en [17].

6.2.1.4.3 Entrenamiento y validación

El entrenamiento del modelo se realizó combinando técnicas de k- cross-validation y early stopping. Respecto a la primera, esta se utiliza en 5 folds de separación 80% y 20% de entrenamiento y validación, respectivamente, con el motivo de combatir el sobreajuste utilizando la naturaleza iterativa de este método. Por otro lado, la segunda se utiliza en 10 rondas y permite controlar el sobreajuste que se pueda generar al entrenar en cada fold.

6.2.1.4.4 Balance de clases y optimización de hiperparámetros

Los hiperparámetros del modelo son los que se componen por defecto en un modelo XGBoost. Dentro de estos, como fue dicho, el que se encarga del desbalance de clases es el scale_pos_weight. En el Anexo W se puede ver una descripción de cada hiperparámetro y parámetro del modelo [29]. Por otro lado, para obtener los hiperparámetros se realiza una optimización de ellos mediante la librería Optuna, maximizando el AUC de validación en un máximo de 100 iteraciones de entrenamiento y validación.

6.2.1.4.5 Reducción de atributos y re-optimización

Acorde a investigaciones y experiencias previas, al ver la importancia de las variables se podrían encontrar características irrelevantes y, por consiguiente, se podrían excluir. En esa línea, reducir variables no significativas en el modelo también puede acelerar su tiempo de ejecución o incluso mejorar su rendimiento predictivo [49]. Al momento de entrenar y validar con los hiperparámetros optimizados, se procede entonces a realizar una reducción del número de atributos. El criterio es eliminar los que tengan una importancia relativa menor a 0.01, similar al tratamiento por desviación estándar. Se eliminan 30 variables a partir de este tratamiento y la data resultante para entrenar, validar y predecir queda con 50 variables. En el Anexo X se especifican estas variables. Posteriormente, se realiza una nueva optimización de hiperparámetros dada esta nueva data, resultando en los valores finales que se pueden apreciar en el Anexo Y.

6.2.1.4.6 Resultados

Al momento de entrenar el modelo se promedian las métricas de AUC de entrenamiento y validación a lo largo de los folds. Los resultados se pueden ver en la Ilustración 5.

```
Average AUC score on train set (cross-validation): 0.7488908432954486
Average AUC score on valid set (cross-validation): 0.6801618553868919
```

Ilustración 5: Métrica AUC promedio en los sets de entrenamiento y validación

Se puede apreciar que, en efecto, los AUC de entrenamiento y validación están alejados a menos del 10%, por lo que se considera satisfactorio en términos de sobreajuste y de lo obtenido en la PoC, en donde la separación también fue menor a dicho porcentaje.

6.2.1.5 Evaluación

6.2.1.5.1 Reentrenamiento

Luego de iteraciones considerando sólo los datos de entrenamiento, se obtuvo que el AUC fuera de muestra decaía más que al reentrenar previamente con todo el set de entrenamiento y validación, luego, se opta por realizar esto antes de predecir en el de test, siguiendo además la premisa de que entrenar con una mayor cantidad de datos puede mejorar la robustez del modelo. Esto se realiza derivando 2 resultados anteriores: Los hiperparámetros óptimos y el promedio de mejores iteraciones en el número de estimadores del modelo en cada fold, arrojado por la técnica early stopping.

6.2.1.5.2 Resultados

Finalmente, se obtienen en la Ilustración 6 los siguientes resultados al aplicar el modelo en el set de test: Valores shap, curva ROC, curva lift y curva precisión-recall.

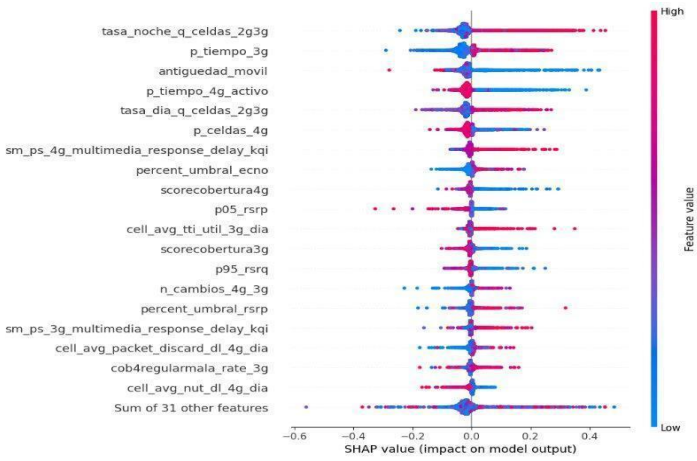


Figura (a): Gráfico de valores shap

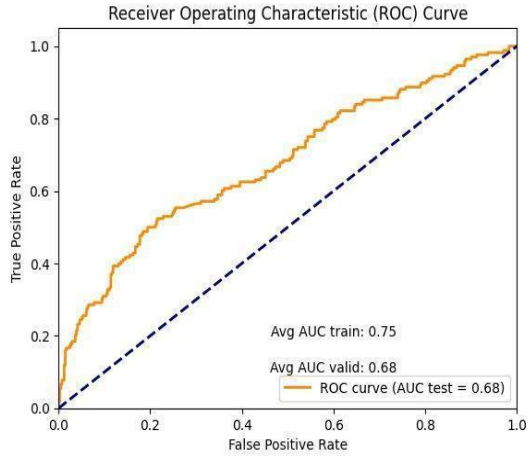


Figura (b): Curva ROC set de test

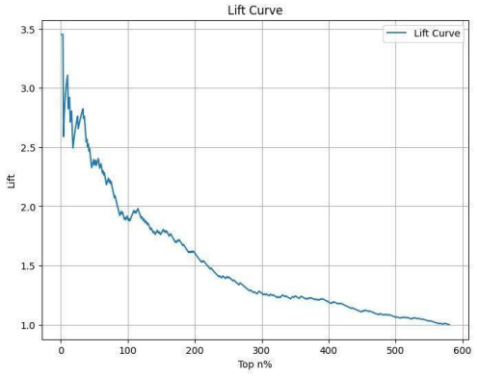


Figura (c): Curva lift

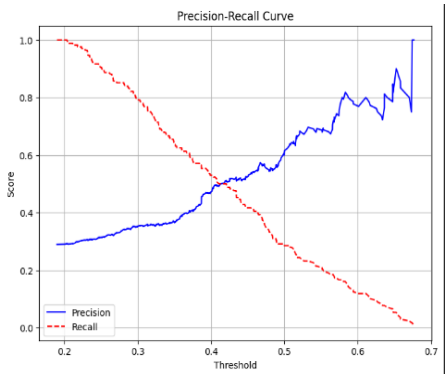


Figura (d) Curva precisión-recall

Ilustración 6: Resultados modelo XGBoost final

En primer lugar, en la figura (a), entre las 20 variables más importantes existen variables

asociadas a todas las tablas incluidas para construir el tablón. Luego, en la figura (b), el AUC de test obtenido es de 0.677, el cual calza virtualmente con el de validación y está dentro del rango obtenido en la PoC). Además, respecto a la figura (c), el modelo predice correctamente a los clientes con mayor propensión a fugarse por red más de 3 veces mejor que un modelo base. Por último, de la figura (d), se desprende que el threshold óptimo es de 0.4128, es decir, a todos los clientes fugados que el modelo le estime una probabilidad igual o mayor a 0.4128, será clasificado como fuga por fallas técnicas de redes (como 1).

6.2.2 Desarrollo de resultados

6.2.2.1 Propensión por cada cliente fugado

Se utiliza el modelo sobre una nueva tabla llamada bi_bt_actividad_comercial, la cual posee a los clientes que hacen port out mes a mes. Para el mes de prueba (diciembre) se obtiene una cantidad de clientes fugados del orden de 50 mil (coincidente con la tasa promedio de fuga mensual entre 1 y 2%). Luego, se realiza el mismo procedimiento de limpieza y procesamiento realizado con los datos de entrenamiento. Finalmente, se estima la probabilidad a fugarse por falla técnica de cada uno de los clientes y se usa el threshold para su atribución en 1 o 0, una descripción de la tabla resultante se da en la Tabla 3:

Tabla 3: Descripción de la tabla de propensión de fuga por red de los clientes fugados

Variable	Descripción
Movil_e (str)	Identificador del cliente
Predict_proba_value (flt)	Probabilidad de propensión de fuga por motivos de red
Predict_value (flt)	Atribución de fuga por red en 1 o 0 usando el threshold

De aquí, se deriva el 1er entregable del modelo (llamado predict_df) desarrollado para el mes de diciembre de 2023 a modo de visualización inicial.

6.2.2.2 Tasas de propensión promedio por POP

Para llegar al 2do entregable, en primer lugar, se sigue el esquema de la Ilustración 7.

Ilustración 7: Procedimiento inicial del 2do entregable

Como consideraciones adicionales se menciona lo siguiente:

- Se decide utilizar como asignación a los POP las celdas 4G más traficadas por los clientes durante el día, esto a partir de conversaciones con el equipo y la contraparte de que este criterio representa a la mayor cantidad de clientes.

- Se considera la tabla calidad_4g_diario para desprender los atributos NUT y PRB de las celdas, siguiendo lo recopilado en la sección 6.2.1.1.

Luego, la tabla final resultante (llamada pop_df), utilizada como el entregable 2 del modelo, se describe a continuación en la Tabla 4:

Tabla 4: Descripción de la tabla de tasa promedio de fuga por red en cada POP

Variable	Descripción
Pop_m_id (str)	Nombre del POP
Lat_wgs84 (flt), lon_wgs84 (flt)	Latitud y longitud del POP
Comuna (str)	Comuna del POP
Prom_prb (flt)	Promedio del prb de la celda asociada a dicho POP
Prom_nut (flt)	Promedio del nut de la celda asociada a dicho POP
Q_movil_e (int)	Cantidad de clientes fugados asociados al POP
Avg_predict_proba_value (flt)	Promedio de propensión de fuga por red del POP

Esta tabla cuenta con un total de 48649 clientes fugados y 6341 POP, aquellos que no tienen clientes fugados asociados tienen un valor nulo en q_movil_e. Además, se considera latitud, longitud y comuna del POP, esto para poder localizarlos.

6.3 Análisis de consistencia y estabilidad en el tiempo del modelo

Para esta parte, se replican los 2 entregables para los meses de noviembre y octubre, emulando el rango de tiempo en que eventualmente se implementará el modelo.

6.3.1 Predicción de fuga por red e intervalos de confianza (IC)

A partir de los promedios mensuales de fuga por red de la encuesta, se construye un IC al 95% para estos 3 meses, utilizando la variable 'churn' (1 si 'motivo' es 'Mala señal o cobertura" y 0 si es 'Otros'). Se consideró una distribución normal para calcular este intervalo, dada la curva vislumbrada en la Ilustración 3. Por otra parte, en el Anexo Z se muestra un vistazo del código desarrollado. En la Ilustración 8 se puede ver el resultado.

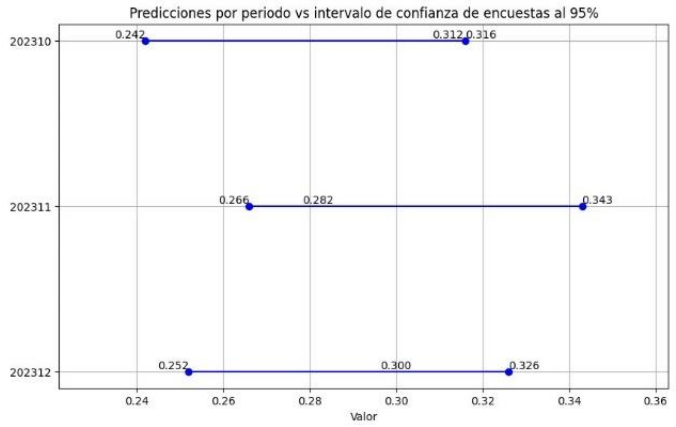


Ilustración 8: Promedio de atribución de fuga por red y su IC para diciembre, octubre y noviembre del 2023

De aquí, se puede ver que el modelo es capaz de mantenerse dentro del IC en cada periodo.

6.3.2 POP coincidentes entre los 100 peores

Este análisis consta de tomar los 100 POP con mayores tasas de propensión promedio de fuga por red y ver si se mantienen en los meses consecutivos considerados. Además, se agrupan los POP en base a cuartiles de la cantidad de clientes fugados asociados, esto para tener más detalle en el análisis respecto a POP críticos. El resultado de este análisis se puede evidenciar en la Ilustración 9.

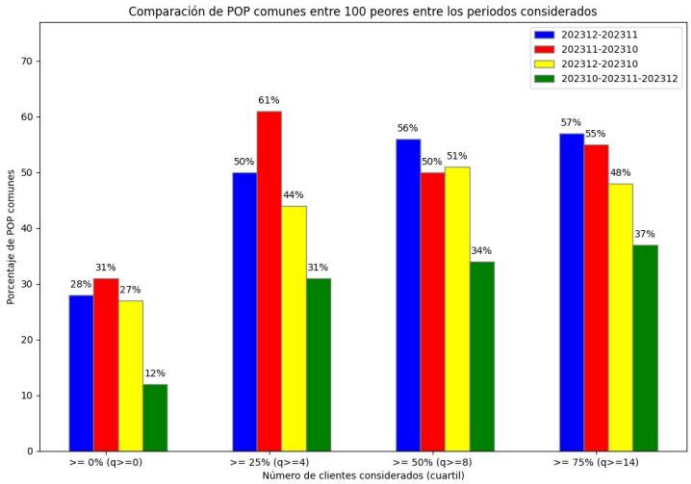
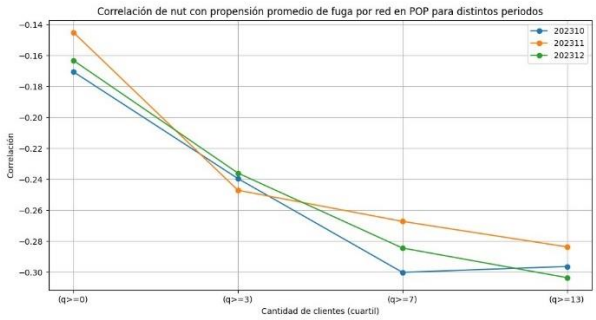


Ilustración 9: Consistencia de POP críticos entre diciembre, octubre y noviembre

De aquí, se puede ver que la cantidad de POP consistentes como críticos es, en general, creciente mientras más clientes asociados, llegando a alcanzar hasta un 60%.

6.3.3 Consistencia de atributos críticos

En esta parte, se calcula la correlación simple entre la tasa promedio de fuga por red y el promedio tanto de NUT como de PRB asociado a los POP. Estos resultados se pueden apreciar en las figuras (a) y (b) de Ilustración 10 para el NUT y PRB, respectivamente.



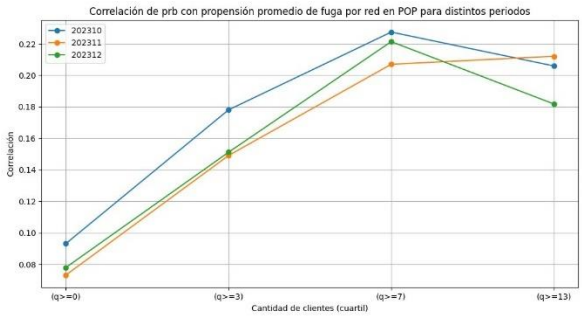


Figura (a): Correlación con NUT

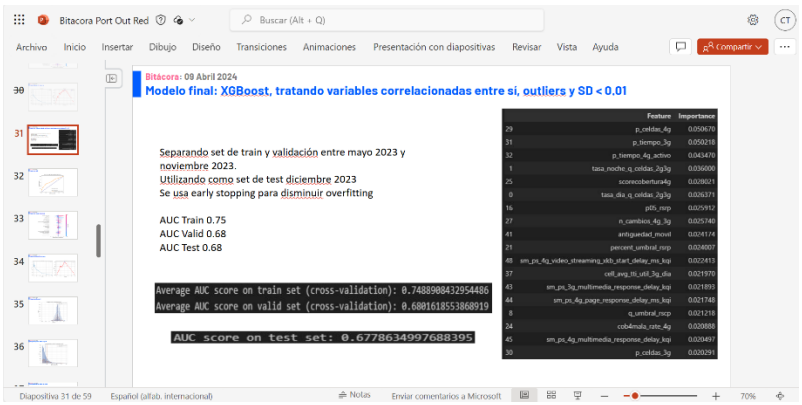
Figura (b): Correlación con PRB

Ilustración 10: Correlación de fuga por red con NUT y PRB por cuartil y en los 3 meses

De aquí se puede ver que, en general, a medida que un POP es más crítico, su velocidad de descarga de datos es menor, mientras que la utilización de sus recursos es mayor.

6.4 Documentación del trabajo

Se muestra en la Ilustración 11 un vistazo de la bitácora del proyecto (referente al modelo final resultante) y la organización de los pipelines de desarrollo en carpetas en VSCode en las figuras (a) y (b), respectivamente. Además, en el Anexo A1 se puede ver un vistazo al repositorio virtual del proyecto, en donde el desarrollo se encuentra respaldado en él.



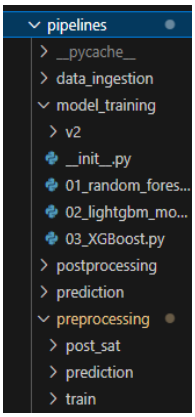


Figura (a): Vistazo bitácora del proyecto

Figura (b): Pipelines de desarrollo

Ilustración 11: Documentación principal del trabajo

A pesar de que el trabajo no involucra un traspaso del código o sesiones de inducción a este con la contraparte (dado que es parte de etapas siguientes), existen comentarios metódicos dentro de los scripts, así como un documento llamado “Readme” dentro del proyecto con un instructivo de cómo configurar su desarrollo para cada usuario. En los Anexos B1 y C1 se pueden ver vistazos de comentarios y del “Readme”, respectivamente.

7. Discusiones y Conclusiones

7.1 Discusiones

7.1.1 Interpretación de resultados e impacto

A partir de los resultados del modelo se puede considerar lo siguiente:

- El modelo final fue un XGBoost, lo que es consistente con estudios previos (como en [17]). Según la Ilustración 6, posee un rendimiento deseable en AUC de 0.68 fuera de muestra que es prácticamente idéntico al del set de validación y está dentro del rango de la PoC. Además, por su curva lift, el modelo predice correctamente a los clientes con mayor propensión a fugarse por red más de 3 veces mejor que un modelo base. En conclusión, se demuestra que el sobreajuste se minimiza y que el modelo es robusto fuera de muestra, logrando predecir correctamente como más probable la atribución de fuga por red en un 68% de probabilidad, permitiendo a la contraparte priorizar las inversiones de mejoras considerando esta premisa.

- Según el gráfico de valores shap, entre las 20 variables más importantes existen variables de todas las tablas, lo que valida sus inclusiones provenientes de la fase de levantamiento. En adición, se ratifica que cuando un cliente navega predominantemente en celdas 3G, tiende a tener una peor experiencia y a fugarse entonces por red, en comparación a lo que ocurre con la tecnología 4G. Luego, en términos del negocio (a partir de lo expuesto en la sección 2.2), una mejora a los POP que con peor experiencia debería ser aumentar la cobertura 4G de sus antenas.

- El modelo es capaz de generar los entregables esperados por el negocio, acorde a las Tablas 3 y 4, y al replicar estos para meses fuera del de prueba.

Continuando, respecto a los análisis de robustez en el tiempo y a la documentación del trabajo se puede comentar lo siguiente:

- El modelo es capaz de ser estable en predecir en un horizonte de 3 meses una atribución promedio de fuga por red que en un 95% de certeza se encuentra en el intervalo de valores de la atribución promedio de los clientes totales fugados. Sin embargo, se es consciente, según la Ilustración 8, que el valor predicho se alejaba un tanto más a los límites de este intervalo a medida que se alejaba del mes de diciembre, por lo que se asume que el modelo tiene espacio de mejora en este sentido, aunque iteraciones posteriores realizadas sobre el modelo final (variando formas de preparar los datos y los hiperparámetros) mostraron que ninguna era capaz de sostenerse en el IC para estos 3 meses como la que se muestra aquí.

- El modelo es consistente en identificar POP críticos de forma continua, alcanzando hasta un 60% de coincidencia (entre 202312 y 202311 para una cantidad de clientes asociados mayor a 4), lo cual es lo esperado según la teoría, en dónde POP que son

más críticos (combinan una mayor cantidad de clientes fugados a la vez que una mayor tasa promedio de propensión de fuga por red) deberían mantenerse como tal en el tiempo. El comportamiento esperado es creciente en ese sentido, no obstante, no siempre se cumple según la Ilustración 9, además, el porcentaje de coincidencias se es consciente que puede ser mayor. Luego, los resultados si bien son consistentes, pueden ser iterados desde la construcción del modelo.

- Finalmente, el modelo es consistente con atributos críticos en cada mes del horizonte. Por el lado del NUT, es esperable que, si la celda 4G más traficada durante el mes previo a la fuga tiende a tener valores bajos de velocidad de descarga de datos, entonces es más probable que exista una mayor propensión promedio de fuga por mala experiencia, y esto se incrementa a medida que se toman POP más críticos. Por el lado del PRB, es esperable que si la celda 4G más traficada durante el mes previo a la fuga tiende a tener una utilización alta (o a estar sobrecargado), entonces es más probable que exista una mayor propensión promedio a fugarse por mala experiencia, dado que esto puede causar intermitencias en la señal, y esto se incrementa en POP más críticos. Ahora bien, se es consciente que estas magnitudes de correlación podrían ser más altas, o también el hecho de que en el 4to cuartil de clientes el comportamiento deja de ser lineal. Luego, al igual que en los puntos anteriores, estos resultados son mejorables desde la construcción del modelo.

- Cerrando con la documentación, existe una bitácora, un desarrollo y un respaldo de este en formatos esperados por la empresa y el negocio (tal como se ve en la Ilustración 11), luego, se cumple con este resultado. No obstante, se deja abierta la necesidad posterior de capacitación del personal para interpretar y actuar sobre los resultados del modelo al momento de implementarlo como producto.

Como conclusión de esta sección, el modelo cumple con sus propósitos planteados desde el objetivo general y los específicos, por lo que está en posición de causar impacto en el negocio, complementando las estrategias de retención de clientes existentes desde un punto de vista de mejoras en infraestructura en caso de ser implementado como producto en las etapas posteriores. No obstante, existe espacio para su mejora antes que se realice esto y de, con ello, armar un plan de priorización de inversión para el año siguiente.

Por último, se considera que este trabajo va a permitir expandir las líneas de investigación no sólo en la industria de las telecomunicaciones, esto se fundamenta en que, desde la sección de marco conceptual y la experiencia, es sabido que es más común encontrar estudios sobre el fenómeno de la fuga a futuro (prediciendo quién se fugará) que desde un punto de vista regresivo (atribución de la misma). Luego, se considera que con este trabajo se deja un marco inicial de metodología y resultados de un nuevo enfoque del fenómeno de la fuga no sólo en la industria de interés, sino que con el potencial de ser considerado para cualquier otra, dado que utiliza modelos y metodologías (como CRISP-DM) que son transversalmente utilizados para estudiar comportamientos de clientes con datos.

7.1.2 Análisis crítico y de mejoras del proyecto

En esta sección, se describirán instancias de mejoras del proyecto desde su ideación y hasta los resultados, esto desde la percepción personal y lo realizado en este informe.

En primer lugar, se asume que existe una instancia de mejora en la sección de levantamiento, en particular, en lo relacionado a la recopilación de las variables y datos disponibles. En ese sentido, en el transcurso del trabajo se dio cuenta que existían otras tablas que no habían sido consideradas desde etapas previas ni tampoco mencionadas desde el equipo, estos datos son referentes al tráfico horario de los clientes, así como la calidad horaria de las celdas (tablas análogas a cdr_bt_kpi_trafico_diario y redes_bt_calidad_3g/4g_diaria, respectivamente). La falta de visualización de estas tablas se debió principalmente a la ausencia de instancias de conversaciones con expertos del proceso, donde se podrían haber levantado todas las tablas relevantes sin depender de las heredadas de la etapa previa. Esto ocurrió porque se priorizó mostrar avances en el desarrollo del modelo debido a los plazos establecidos. Una forma de revertir esta situación es estalecer, al inicio del proyecto y de acuerdo con la contraparte, un plazo para mostrar avances que incluya generar las intancias necesarias de conversaciones.

Para futuras iteraciones de este modelo se considera que incluir estas tablas podría mejorar la precisión del modelo fuera de muestra y, con ello, su impacto en la empresa y en el negocio. Además, con ello, podrían mejorar los resultados en la sección 6.3.

Otro espacio de mejora derivado de la metodología resulta en la etapa de preparación de la data del CRISP-DM. En ese sentido, se deja en claro que los métodos derivados para este informe se basaron en 3 aristas: Recomendaciones del equipo respecto a experiencias previas, experiencia personal e investigaciones generales. No obstante, se es consciente que las justificaciones para aplicar uno u otro tratamiento pueden derivar no sólo de estas 3 aristas, sino, por ejemplo, de una arista relacionada a considerar cuáles tratamientos podrían ser más útiles o no necesarias dependiendo del modelo que se utiliza, una arista relacionada más a proyectos similares que se hayan hecho o una arista relacionada a usar una mayor variedad de modelos. La no inclusión de estas aristas se debe a los plazos y tiempos acotados acordados con la contraparte de entrega de resultados, luego, la recomendación en esta parte es definir al inicio del proyecto un mayor tiempo a realizar investigaciones científicas que respalden las decisiones tomadas y su impacto.

Una consideración de lo expuesto en este último párrafo en una futura iteración del modelo o en trabajos con un objetivo similar en otra organización o industria podría ocasionar resultados más precisos en la evaluación del modelo en términos de rendimiento y sobreajuste, así como de los análisis de consistencia realizados.

Todas estas instancias de mejoras se pueden relacionar con el hecho de que el trabajo desarrollado es parte de una etapa intermedia dentro de lo que comprende un proyecto de datos en la empresa. En ese sentido, no se fue partícipe de las instancias de ideación del problema, del proyecto y de los objetivos ni con el equipo ni con la contraparte asociada, luego, desde el inicio existió una desconexión respecto a las necesidades y al impacto en el negocio, lo cual además fue puesto en segundo plano debido a la necesidad de cumplir con los plazos establecidos. Además, el no involucramiento en fases posteriores puede también limitar la visión global respecto al proyecto y su verdadero impacto en una implementación, lo cual pudo mejorar, por ejemplo, el mismo proceso de levantamiento.

Como complemento, puede existir un espacio de mejora en los supuestos tomados para una posterior revisión del caso de negocio e impacto real en la fuga. Esto tiene que ver con 2 aspectos: En primer lugar, es necesario analizar si el 30% de fuga por red de la encuesta

puede representar a la población de clientes fugados, dado que las respuestas entregadas podrían, por ejemplo, estar ligadas a factores ajenos a la percepción de la experiencia. En segundo lugar, es importante evaluar si el criterio de asignación de clientes a los POP por mayor tiempo de navegación 4G representa la experiencia previa a la fuga. Esto se debe a que podrían estarse priorizando POP que no son los que presentan las fallas técnicas.

En esa misma línea, se considera relevante revisar los siguientes aspectos en pos de brindarle una mayor robustez a los resultados del modelo ante cambios en las variables: Evaluación en otros datos de test y aplicación de otras técnicas de balance de clases y representatividad de la asignación de fuga a los POP. Respecto al 1er punto, esto no se consideró en el trabajo debido a que no se contaban con datos de encuestas actualizados al momento de evaluar el modelo. En relación con el 2do punto, se optó por balancear las clases desde un punto de vista de optimización de hiperparámetros por consisderarlo un método más integral a cada uno de los modelos, no obstante, se es consciente que existen otras técnicas en la data science para ello, por lo que resulta interesante probarlas.

Como conclusión de esta sección, el proyecto es mejorable tomando los siguientes aprendizajes para posibles aplicaciones o iteraciones de este estudio en futuros trabajos:

- Es relevante dedicar el tiempo para generar todas las instancias necesarias de conversaciones con expertos del proceso de priorización, esto para realizar un levantamiento y exploración de datos relevantes de manera completa, más allá de lo que se haya realizado o recomendado previamente desde el equipo o el negocio.

- Además, es relevante tener un respaldo científico y realizar un estudio exhaustivo de cómo funcionan los algoritmos de machine learning y cuál es la mejor forma de preparar los datos en base a estos puede permitir librarse de posibles ambigüedades o faltas de justificación más científica en el trabajo, lo que a su vez potencialmente deriva en resultados más robustos y sostenibles en el tiempo.

- En adición, se destaca lo relevante que es mantener una comunicación constante (y directa) con la contraparte involucrada no sólo en un proyecto con datos, si no que en cualquier rubro. Esta relevancia viene de la mano con entender de primera fuente sus necesidades y el impacto para la empresa, procurando así que todas las fases posteriores de trabajo se realicen más rigurosamente y se permita, con ello, volver a tiempo sobre puntos que no se hayan realizado de manera satisfactoria.

- Finalmente, se comenta el potencial impacto, en una posterior revisión del caso de negocio al momento de implementar, de revisar el carácter representativo en la experiencia de la encuesta utilizada y del criterio de asignación de los clientes a los POP. Además, se reflexiona sobre aspectos como la evaluación del modelo en datos nuevos y la aplicación de más técnicas de balance de clases.

7.2 Conclusiones

El objetivo principal de este proyecto era desarrollar un modelo de atribución de la fuga a fallas técnicas de redes y asignarla a los POP, garantizando su estabilidad y consistencia en el tiempo para una posible implementación como producto. Los resultados muestran

que se ha cumplido este objetivo, ya que el modelo proporciona una herramienta precisa para identificar y asignar a los POP clientes con esta atribución de fuga, apoyando el trabajo de la contraparte SmartCapex en decidir en cuales POP invertir al año siguiente.

En ese sentido, conectando con los objetivos, se realizó un levantamiento acorde a conocer la etapa del proyecto que se iba a abordar en el trabajo, los conceptos y herramientas necesarios para su desarrollo, y los insumos provenientes de las etapas anteriores. Luego, esta parte logra rescatar el objetivo del proyecto, tablas de etapas previas y recomendaciones de otras, el horizonte de tiempo a considerar y los resultados que el negocio espera. No obstante, se considera que los resultados de este objetivo son mejorables para futuras iteraciones u otros trabajos que requieran un levantamiento, esto debido a que no se dio cuenta de la existencia de todas las tablas que estaban disponibles.

Respecto a la construcción del modelo de predicción, los resultados indican que se logra encontrar un modelo XGBoost capaz de poseer un AUC prácticamente idéntico tanto dentro como fuera de muestra, esto gracias principalmente a la utilización de técnicas de entrenamiento como validación cruzada y eary stopping, y a la optimización de hiperparámetros, los cuales trabajando en conjunto minimizaron el sobreajuste, a la vez que maximizaron el AUC. Además, este modelo mantiene el rendimiento obtenido en la PoC, lo que hace este trabajo consistente con lo ya realizado. En adición, se comprueba que, en general, los clientes que más tiempo navegan en 3G tienen una mayor probabilidad a fugarse por red en comparación con los que lo hacen en 4G. El hecho de que clientes naveguen en mayor proporción en celdas con tecnología 3G puede deberse a que en dichas zonas exista una baja cobertura 4G, por lo que se podría hacer necesario invertir en aumentar dicha tecnología. Ahora bien, el cumplimiento de este objetivo es mejorable en el sentido del respaldo científico utilizado en los tratamientos de los datos y de la investigación realizada sobre cómo cada modelo se puede ver beneficiado de estos.

Respecto a la generación de los entregables, el modelo logra desarrollarlos para periodos incluso fuera del utilizado en su evaluación. No obstante, este desarrollo se pudo haber beneficiado de un mejor levantamiento de tablas en la primera fase o de un mayor respaldo científico respecto al criterio utilizado para asignar a los clientes a los POP.

Finalmente, en lo que respecta a los análisis de estabilidad y documentación, el modelo es capaz de ser consistente y coherente en el tiempo respecto a las predicciones de la propensión promedio a fugarse por red, consistencia de POP críticos y coherencia en el comportamiento de atributos relevantes para el negocio. Sin embargo, un mayor respaldo en la literatura o una consideración de todas las limitantes anteriores pueden mejorar los resultados obtenidos en este objetivo en términos de precisión. En adición, la documentación del desarrollo del proyecto se realizó en estándares de la empresa y el negocio para su entendimiento, luego, se cumple con este apartado en el objetivo.

Como conclusión final, queda el aprendizaje para futuros trabajos (o iteraciones del presente) de que un proyecto, para ser realizado de forma óptima y autónoma, requiere no solo de un involucramiento total y global, respondiendo ante la contraparte de manera directa y no indirecta o como apoyo, sino también de una planificación rigurosa de los tiempos y plazos dedicados a la investigación conversación con expertos, y desarrollo en base no sólo a las necesidades del negocio, sino que también a la relevancia de cada etapa que involucra, en este caso, un proyecto de datos con posterior modelamiento.

Bibliografía

[1] Bnamericas. (16 de abril de 2024). La situación del mercado de telecomunicaciones de Chile. La situación del mercado de telecomunicaciones de Chile - BNamericas

[2] Ministerio de Transportes y Telecomunicaciones, Subsecretaría de Telecomunicaciones. (2024). Sector Telecomunicaciones Cuarto Trimestre 2023. Informe-telecomunicaciones-Dic23.pdf (subtel.gob.cl)

[3] Entel S.A. (2024). Memoria Integrada 2023. Memoria Entel 2023.pdf (modyocdn.com)

[4] Aravena, S. (12 de abril de 2024). Presidente de Entel alerta de importantes desafíos de rentabilidad que enfrentan las empresas de telecomunicaciones. La Tercera. Presidente de Entel alerta de importantes desafíos de rentabilidad - La Tercera

[5] Entel S.A. (2024). Llevamos 60 años acercando las infinitas posibilidades que da la tecnología. Entel: Líder en Tecnología y Telecomunicaciones

[6] Entel S.A. (2024). Ejecutivos. Entel: Líder en Tecnología y Telecomunicaciones

[7] Entel S.A. (2024). 4Q 2023 RESULTS. 4Q 2023 Results Press Release.pdf (modyocdn.com)

[8] Entel S.A. (2024). 4Q23 Results Presentation [17]. PPT_4Q23_Versio_n_pweb_1_.pdf (modyocdn.com)

[9] Miranda, M. (11 de marzo de 2024). Entel gana premio a la mejor red móvil de Chile en MWC24. La Hora. Entel gana premio a la mejor red móvil de Chile en MWC24 (lahora.cl)

[10] Entel S.A. (2024). 4Q23 Results Presentation [9]. PPT_4Q23_Versio_n_pweb_1_.pdf (modyocdn.com)

[11] Entel S.A. (2024). Análisis Razonado de los Estados Financieros Consolidados al 31 de diciembre del 2023. Ana_lisis_Razonado92580000_202312.pdf (modyocdn.com)

[12] C. B. Bhattacharya. (1998). “When customers are members: customer retention in paid membership contexts,” Journal of the Academy of Marketing Science.

[13] Entel S.A. (2023). Memoria Integrada 2022. 230419_Entel_Memoria_2023_Libro_interactiva.pdf (modyocdn.com)

[14] Amazon Web Services. (2023). ¿Qué es el aprendizaje automático?. ¿Qué es el machine learning? - Explicación sobre el machine learning empresarial - AWS (amazon.com)

[15] Gamco. (2021). Modelo predictivo. Concepto y definiciones. Qué es Modelo predictivo Concepto y definición. Glosario (gamco.es)

[16] Keyrus. Las 11 técnicas más utilizadas en el modelado de análisis predictivo. Las 11 técnicas más utilizadas en el modelado de análisis predictivos - Insight | Keyrus

[17] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine in big data Customer churn prediction in telecom using machine learning in big data platform | Journal of Big Data (springer.com)

[18] Máxima formación. (2020). Qué son los árboles de decisión y para qué sirven. Qué son los árboles de decisión y para qué sirven - Máxima Formación (maximaformacion.es)

[19] Data Scientist. (25 de enero de 2024). Random Forest: Bosque aleatorio. Definición y funcionamiento. Random Forest: Bosque aleatorio. Definición y funcionamiento (datascientest.com)

[20] Valencia, L. Qué es Gradient Boosting. Compartimoss. Introducción a Gradient Boosting y su implementación con Microsoft LightGBM | CompartiMOSS

[21] InteractiveChaos. Entrenamiento de modelos. Entrenamiento de modelos | Interactive Chaos

[22] Shaip. (1 de marzo de 2022). ¿Cuál es el volumen óptimo de datos de entrenamiento que necesita para un proyecto de IA? ¿Cuál es el volumen óptimo de datos de entrenamiento que necesita para un proyecto de IA? | Shaip

[23] IBM. ¿Qué es sobreajuste? ¿Qué es el sobreajuste? | IBM

[24] Microsoft Learn. (1 de junio de 2023). Modelo de validación cruzada. Modelo de validación cruzada: referencia del componente - Azure Machine Learning | Microsoft Learn

[25] Sotaquirá, M. (31 de julio de 2023). Validación cruzada y k-fold cross validation. Codificandobits. Validación cruzada y k-fold cross-validation | Codificando Bits

[26] Meza, A.; Chue, J. 2020. Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil. Natura@economía 5(2). Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil - Dialnet (unirioja.es)

[27] Fernández, A.; Río, S.; Chawla, N.; Herrera, F. 2017. An insight into imbalanced Big Data classification: outcomes and challenges. Complex & Intelligent Systems 3. An insight into imbalanced Big Data classification: outcomes and challenges | Complex & Intelligent Systems (springer.com)

[28] Interactive Chaos. Otros parámetros. Otros parámetros | Interactive Chaos

[29] Amazon Web Services. (2024). Hiperparámetros de XGBoost. Hiperparámetros de XGBoost - Amazon SageMaker

[30] Gallardo Burns, C.A. (2021). MODELO DE PREDICCIÓN DE FUGA DE CLIENTES EN EMPRESA SAAS DE INTELIGENCIA LOGÍSTICA [Memoria para optar al título de Ingeniería Civil Industrial, Universidad De Chile]. Modelo-de-prediccion-de-fuga-de-clientes-en-empresa-SaaS-de-inteligencia-logistica.pdf

[31] Brainfood. Métricas aplicadas en modelos de propensión. Métricas aplicadas en modelos de propensión - BrainFood

[32] All Awan, A. (marzo de 2024). Una introducción a los valores Shap y a la intepretabilidad del machine learning. DataCamp.

Una introducción a los valores SHAP y a la interpretabilidad del machine learning | DataCamp

[33] Trevisan, V. (17 de enero de 2022). Using SHAP Values to Explain How Your Machine Learning Model Works. Medium. Using SHAP Values to Explain How Your Machine Learning Model Works | by Vinícius Trevisan | Towards Data Science

[34] HelpPortal. La Curva Lift. La curva lift | SAP Help Portal

[35] MicrosoftBuild. (7 de junio de 2023). Ajuste de hiperparámetros de un modelo (v2). Ajuste de hiperparámetros de un modelo (v2) - Azure Machine Learning | Microsoft Learn

[36] Ciberseguridad. ¿Qué es optuna? Hiperparámetros, enfoque y características. ¿Qué es Optuna? Hiperparámetros, enfoque y características (ciberseguridad.com)

[37] Saavedra, R. (26 de agosto de 2022). Búsqueda eficiente de hiperparámetros con Optuna + Sklearn para XGBoost y lightGBM. Medium. Búsqueda eficiente de hiperparámetros con Optuna + Sklearn para XGBoost y LightGBM | by Rodrigo Saavedra | Rappi Tech

[38] Candia B, Caiozzi A. (2005). Intervalos de Confianza. Scielo. Revista Médica de Chile. Intervalos de Confianza (scielo.cl)

[39] WebDesing. (2023). Ventajas y Desventajas de Visual Studio Code 2023: ¿Es la herramienta adecuada para ti? Descubre las ventajas y desventajas de Visual Studio Code en 2023 (webdesigncusco.com)

[40] Wirth, R. & Hipp, Jochen. (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.

[41] Roberto C. (31 de mayo de 2022). CRISP-DM: Las 6 etapas de la metodología del futuro. MBA Usp/Esalq. Crisp-DM: las 6 etapas de la metodología del futuro - Blog MBA Esalq USP (mbauspesalq.com)

[42] Ipmoguide. (2024). CRISP-DM, Metodología de datos. CRISP-DM, Metodología de Datos - iPMOGuide

[43] Luis J. (2023). Entendiendo el sobreajuste en los modelos de machine learning. Medium. Entendiendo el Sobreajuste en los Modelos de Machine Learning | by Jorge Luis | Medium

[44] Ubilla Sababa, N. V. (2022). MODELOS DE PROPENSIÓN DE FUGA Y RELACIÓN DE LAS INTERACCIONES CON CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES. [Memoria para optar al Título de Ingeniería Civil Industrial, Universidad de Chile]. Modelos-de-propension-de-fuga-y-relacion-de-las-interacciones-con-clientes.pdf

[45] Qlik Help. Imputación de nulos. Imputación de nulos | Qlik Cloud Ayuda

[46] Chaud M. (8 de agosto de 2023). Reducción de dimensionalidad. Medium. Reducción de dimensionalidad. Antes de entrenar modelos de ML, una de... | by Matiaschaud | Medium

[47] Gonzalez L. (30 de abril de 2020). Reducción de la dimensionalidad. Aprende IA.

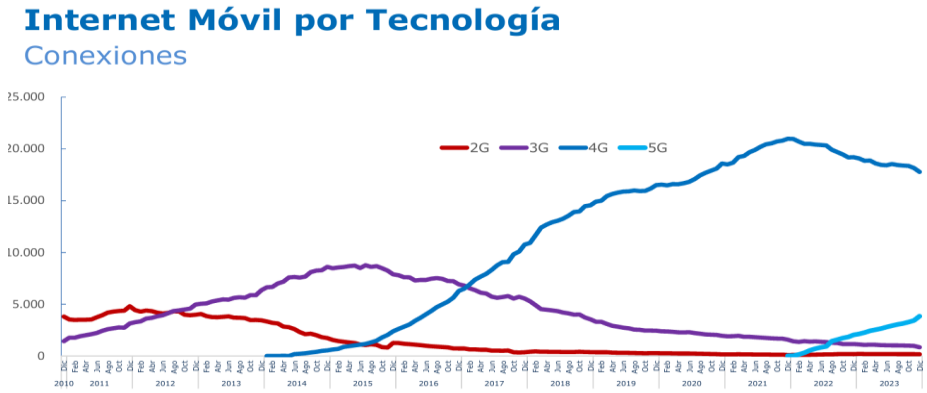
Reducción de la Dimensionalidad - Aprende IA

[48] Aprende Machine Learning. (3 de marzo de 2020). Sets de entrenamiento, test y validación. Sets de Entrenamiento, Test y Validación | Aprende Machine Learning

[49] Zvornicanin E. (18 de marzo de 2024). What is feature importance in machine learning. Baeldung. What Is Feature Importance in Machine Learning? | Baeldung on Computer Science

Anexos

Anexo A. Evolución del uso de las distintas tecnologías en el tiempo (obtenido de [2])



Anexo B. Participación de mercado en conexiones y abonados móviles al cierre del 2022 y 2023 (obtenido de [2])

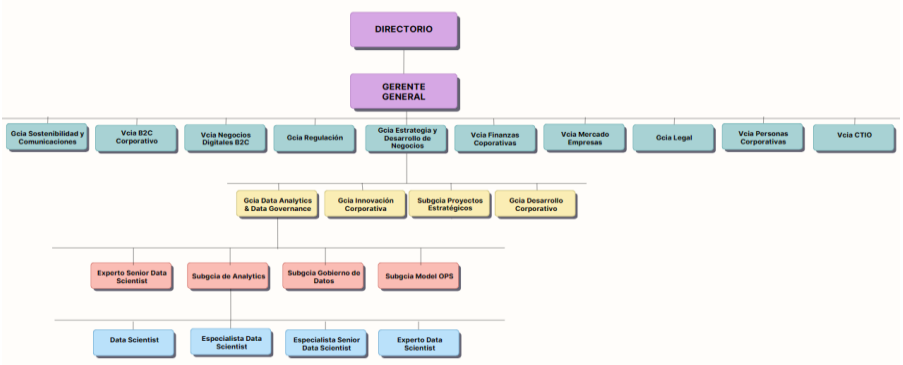
%Particip. Conexiones 3G+4G+5G	Dic 22	Dic 23
Movistar	21,2%	20,9%
Grupo Entel	34,7%	34,6%
Claro	16,9%	16,7%
Wom	25,4%	25,7%
VTR Móvil	1,2%	1,3%
Virgin	0,3%	0,2%
Otros	0,3%	0,6%

Anexo (a): Participación en conexiones

Participación de mercado	Dic 22	Dic 23
Movistar	25,7%	26,6%
ENTEL	32,9%	32,0%
Claro	18,3%	17,9%
Virgin	0,3%	0,2%
WOM	21,6%	21,6%
VTR	1,0%	1,1%
Otros	0,3%	0,5%

Anexo (b): Participación en abonados

Anexo C. Organigrama



Fuente: Elaboración propia

Anexo D. Análisis FODA de Entel

Fortalezas	-Un alto reconocimiento de marca -Negocio altamente diversificado -Líder en participación de mercado y conexiones
Oportunidades	-Demanda cada vez más creciente -Aparición de nuevas tecnologías (como 5G)
Debilidades	-Precios elevados frente a la competencia -Altos costos fijos en infraestructura de redes -Aumento de la fuga anual
Amenazas	-Inestabilidad económica -Competencia creciente -Demanda cada vez más exigente en calidad de la señal

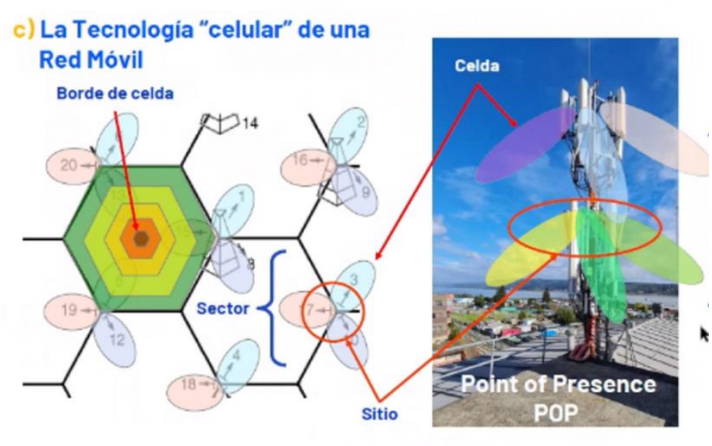
Fuente: Elaboración propia a partir de la información en el capítulo 1

Anexo E. Tasas de abandono (fuga) Chile 4Q 2023 vs 2022 (obtenido en [8])

Enmarcado en negrito (blended churn %) se aprecian los valores combinados. Luego, en "Churn Postpaid", se aprecian los valore del segmento pospago.

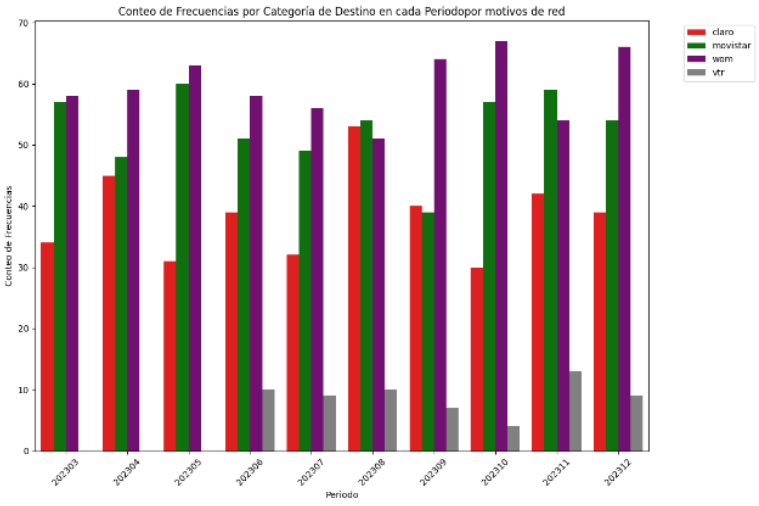
Mobile Customers (th.)	4Q23	4Q22	% Change	Abs. Change	3Q23	% Change
Postpaid Voice	6.061	5.889	2,9%	172	5.971	1,5%
Prepaid Voice	3.294	3.483	(5,4%)	(189)	3.484	(5,4%)
M2M & IOT	1.239	1.048	18,2%	190	1.202	3,1%
Total Mobile Customers (th.)	**10.594**	**10.421**	**1,7%**	**173**	**10.657**	**(0,6%)**
ARPU (CLP)	**6.839**	**7.032**	**(2,8%)**	**(194)**	**6.844**	**(0,1%)**
ARPU Postpaid	9.498	9.850	(3,6%)	(351)	9.602	(1,1%)
ARPU Prepaid	1.168	1.392	(16,1%)	(224)	1.134	3,1%
Blended Churn %	**2,7%**	**1,9%**	**45,0%**	**0**	**1,7%**	**55,8%**
Churn Postpaid	1,86%	1,63%	14,1%	0,2%	1,88%	(1,0%)
Churn Prepaid	4,47%	2,32%	92,3%	2,1%	1,42%	213,9%
Blended MOU	250,4	275,4	(9,1%)	(25)	249,0	0,5%
Postpaid Effective GOU	27,71	22,91	21,0%	5	28,2	(1,7%)
4G Users (th.)	6.911	6.466	6,9%	445	7.238	(4,5%)
5G Users (Th.)	1.517	905	67,7%	612	1.407	7,8%

Anexo F. Esquema general POP-celda-antena (antena son los pilares blancos del POP)



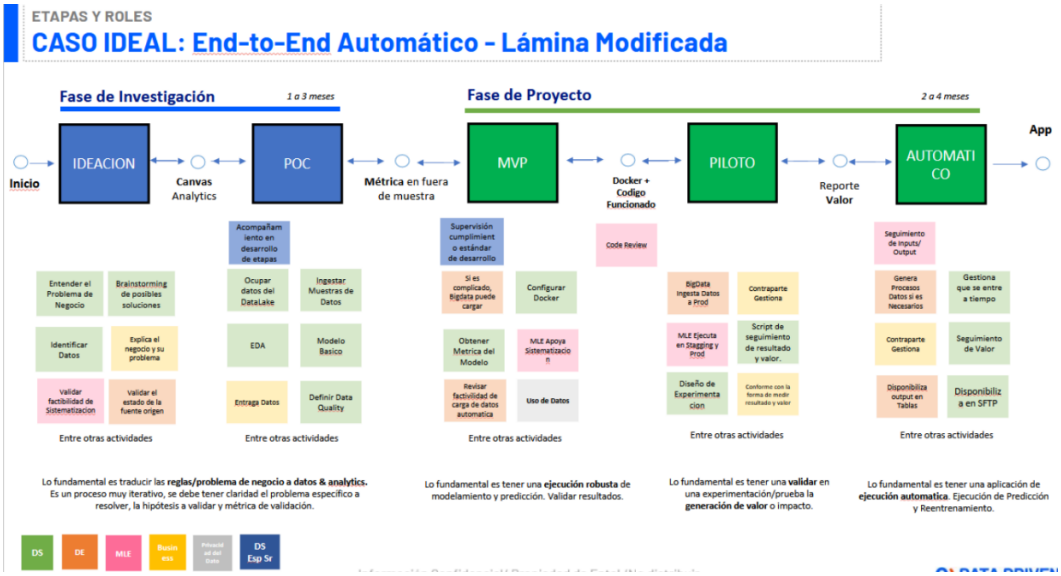
Fuente: Elaboración de Entel

Anexo G. Histograma por mes del destino de los clientes encuestados y fugados por red



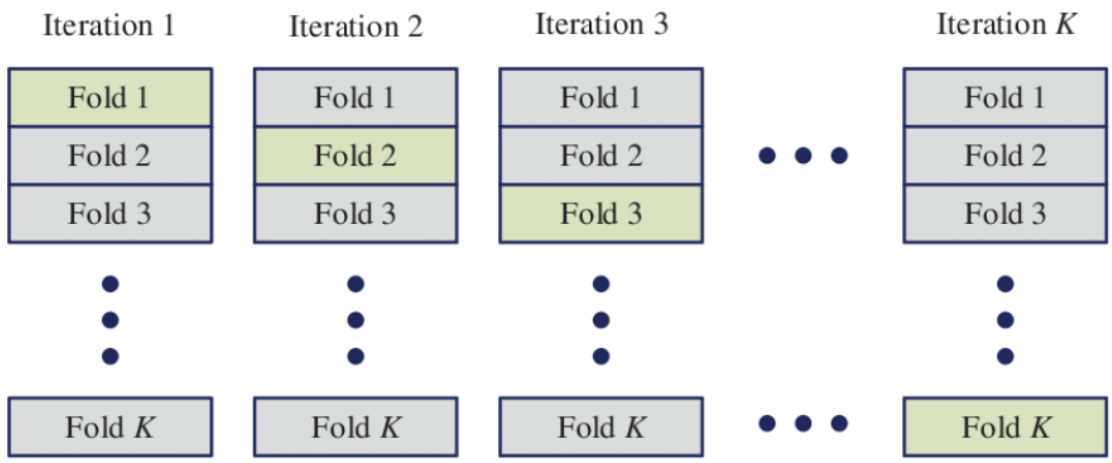
Fuente: Data port_out_survey

Anexo H. Etapas de un proyecto de ciencia de datos en Entel



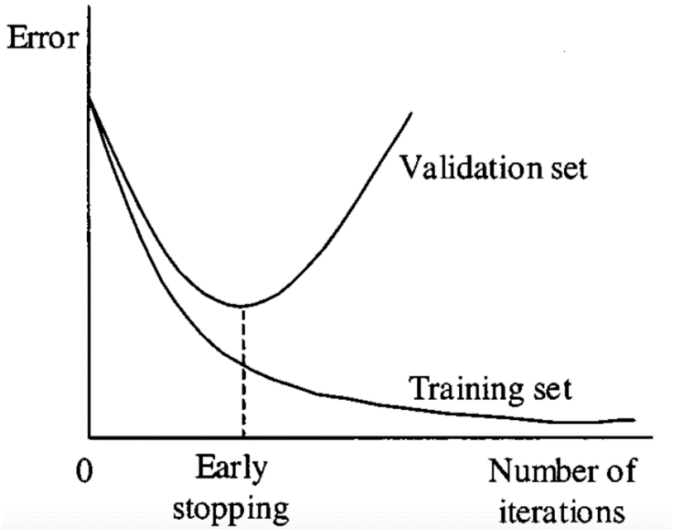
Fuente: Portal de inducción del equipo de Data Science

Anexo I. Esquema resumen de técnica k-fold cross validation utilizada



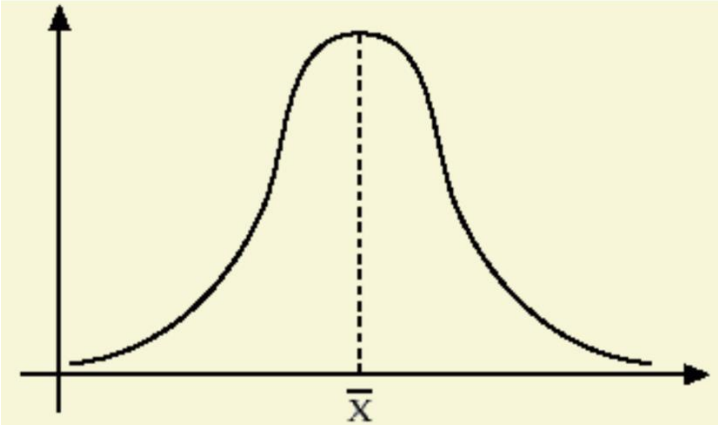
Fuente: ResearchGate

Anexo J. Esquema de cómo funciona Early Stopping en los sets de entrenamiento y validación



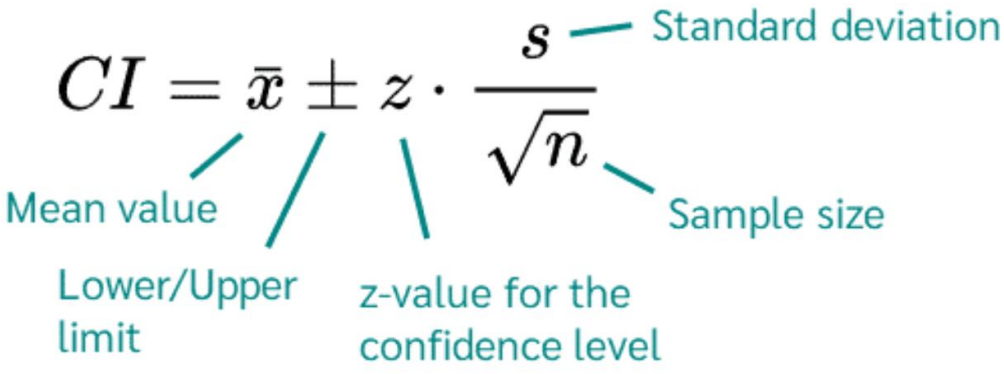
Fuente: ResearchGate

Anexo K. Curva de distribución normal típica



Fuente: Hiru.eus, Distribución Normal

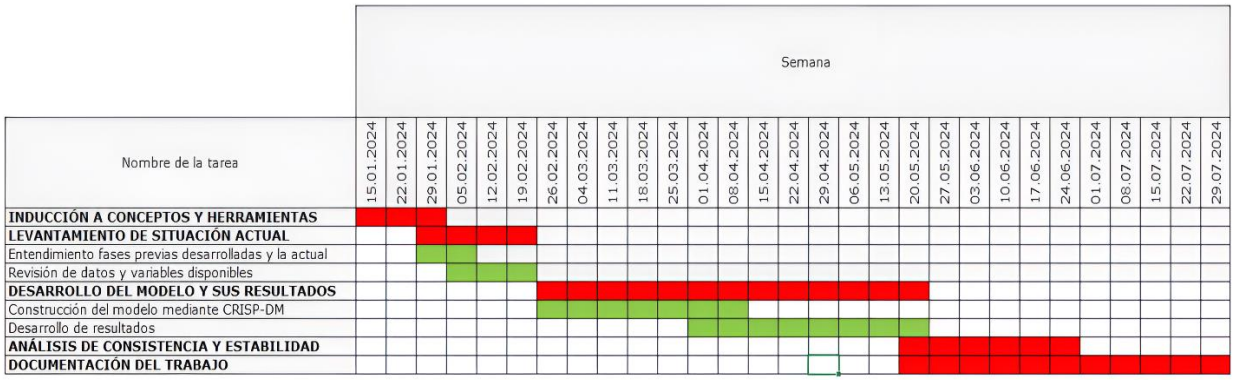
Anexo L. Fórmula para sacar el intervalo de confianza para una distribución normal



En este caso, para un 95% de confianza el valor de z es 1.96. Fuente: DATAtab, 2024, Intervalo de Confianza.

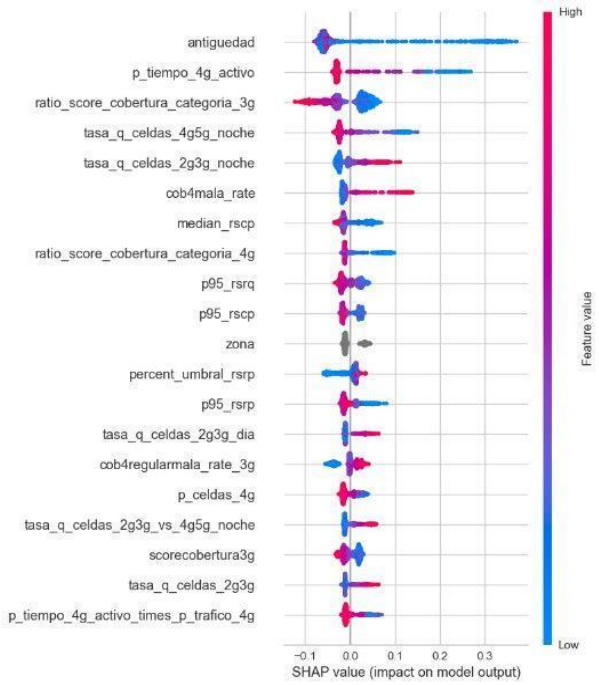
Anexo M. Carta Gantt del proyecto

Comienza el día 15 de julio del 2024 y termina el 31 de julio del mismo año. Además, se considera la etapa de inducción a la empresa, antes de comenzar con la fase 1.



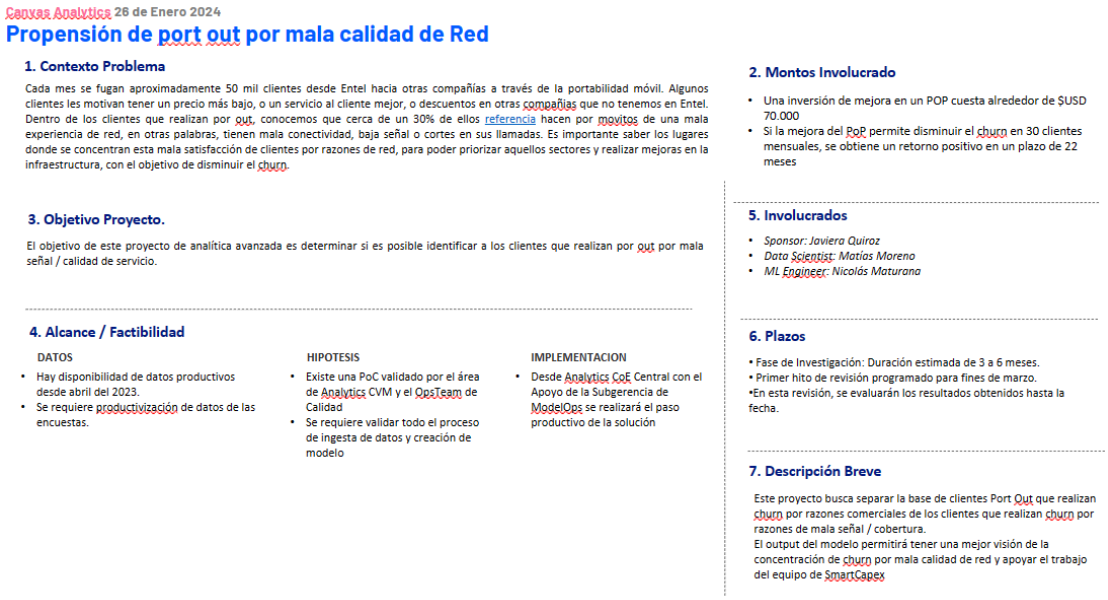
Fuente: Elaboración propia

Anexo N. Valores shap obtenido en la fase de PoC



Fuente: Bitácora de la fase de PoC

Anexo O. Lámina resumen de fases previas desarrolladas, la actual y el objetivo



Canvas Analytics 26 de Enero 2024

Propensión de port out por mala calidad de Red

1. Contexto Problema

Cada mes se fugan aproximadamente 50 mil clientes desde Entel hacia otras compañías a través de la portabilidad móvil. Algunos clientes les motivan tener un precio más bajo, o un servicio al cliente mejor, o descuentos en otras compañías que no tenemos en Entel. Dentro de los clientes que realizan por out, conocemos que cerca de un 30% de ellos referencia hacen por movitos de una mala experiencia de red, en otras palabras, tienen mala conectividad, baja señal o cortes en sus llamadas. Es importante saber los lugares donde se concentran esta mala satisfacción de clientes por razones de red, para poder priorizar aquellos sectores y realizar mejoras en la infraestructura, con el objetivo de disminuir el churn.

2. Montos Involucrado

- Una inversión de mejora en un POP cuesta alrededor de $USD 70.000
- Si la mejora del PoP permite disminuir el churn en 30 clientes mensuales, se obtiene un retorno positivo en un plazo de 22 meses

3. Objetivo Proyecto.

El objetivo de este proyecto de analítica avanzada es determinar si es posible identificar a los clientes que realizan por out por mala señal / calidad de servicio.

4. Alcance / Factibilidad

DATOS
- Hay disponibilidad de datos productivos desde abril del 2023.
- Se requiere productivización de datos de las encuestas.

HIPOTESIS
- Existe una PoC validado por el área de Analytics CVM y el OpsTeam de Calidad
- Se requiere validar todo el proceso de ingesta de datos y creación de modelo

IMPLEMENTACION
- Desde Analytics CoE Central con el Apoyo de la Subgerencia de ModelOps se realizará el paso productivo de la solución

5. Involucrados

- *Sponsor: Javiera Quiroz*
- *Data Scientist: Matías Moreno*
- *ML Engineer: Nicolás Maturana*

6. Plazos

• Fase de Investigación: Duración estimada de 3 a 6 meses.
• Primer hito de revisión programado para fines de marzo.
•En esta revisión, se evaluarán los resultados obtenidos hasta la fecha.

7. Descripción Breve

Este proyecto busca separar la base de clientes Port Out que realizan churn por razones comerciales de los clientes que realizan churn por razones de mala señal / cobertura.
El output del modelo permitirá tener una mejor visión de la concentración de churn por mala calidad de red y apoyar el trabajo del equipo de SmartCapex

Fuente: Bitácora actual del proyecto

Anexo P. Vistazo data port_out_survey, en donde ‘motivo’ es la variable a predecir

Cada cliente está indexado por su número telefónico anonimizado (‘movil_e’).

survey

✓ 0.0s

	destino	fecha	satisfaccion	movil_e	motivo
0	wom	2023-03-01	5	6129bfd3b9ccee1a5b05de9b073d3a157a4bbea3dd6117...	Otros
1	movistar	2023-03-01	5	fd1ce8dabe2b221a56f7de3a18b117e5c5f8cd717f694f...	Otros
2	claro	2023-03-01	5	9b20eeef2744e9fc4040b85411e7ec66cb8c5ac2186447...	Mala señal/cobertura
3	claro	2023-03-01	7	525a86d5b0ae3abd14cfdec63e32e756e154eace6b7469...	Otros
4	claro	2023-03-01	3	acc7d75b864f9bfd04309b5f0f3b4dd15c5f514848a2d1...	Otros
...	...	...	...	...	...
5592	movistar	2023-12-01	6	5eeccc6242e0990e9f102a6277732a9ad6931c700dfde2...	Otros
5593	claro	2023-12-01	7	3bb9c08b18d7954c4db2871103d4d1fc7fa7504e03fda2...	Otros
5594	claro	2023-12-01	5	10c902839e8425c224a9416b03e04ae738b2f39d34b3de...	Mala señal/cobertura
5595	wom	2023-12-01	5	da148599bf53bfb6960f9d7ad99152840942dbeac020f8...	Otros
5596	movistar	2023-12-01	6	75a9927becc4faee527bf694f9cc23a39ae8dff067395f...	Mala señal/cobertura

5597 rows × 5 columns

Anexo Q. Variables de cada tabla y su descripción

bt_ecno_rscp	-periodo: fecha en formato MesAño (entero o int) -movil_e: número telefónico anonimizado (string o str) -median_ecno: mediana de la calidad de la señal 3G recibida en el mes (float o flt) -p05_ecno: percentil 5 de la calidad 3g recibida en el mes (flt) -p95_ecno: percentil 95 de la calidad 3g recibida en el mes (flt) -median_rscp: mediana de la potencia recibida en señal 3g en el mes (flt) -p05_rscp: percentil 5 de la potencia 3g recibida en el mes (flt) -p95_rscp: percentil 95 de la potencia 3g recibida en el mes (flt) -q_umbral_ecno: cantidad de sesiones 3G en el mes bajo un umbral de calidad (int) -q_umbral_rscp: cantidad de sesiones 3G en el mes bajo un umbral de potencia (int) -number_of_sessiones_3g: número de sesiones 3G en el mes (int) -percent_umbral_ecno: porcentaje de sesiones 3G en el mes bajo un umbral de calidad (flt) - percent_umbral_ecno: porcentaje de sesiones 3G en el mes bajo un umbral de potencia (int) -cob4buena_rate: porcentaje de sesiones 3G buenas sobre el total en el mes (flt) -cob4regularbuena_rate: porcentaje de sesiones 3G buenas y regulares sobre el total en el mes (flt) -cob4regular_rate: porcentaje de sesiones 3G regulares sobre el total en el mes (flt) -cob4regularmala_rate: porcentaje de sesiones 3G regulares y malas sobre el total en el mes (flt) -cob4mala_rate: porcentaje de sesiones 3G malas sobre el total en el mes (flt) -scorecobertura3g: score de 1-5 de la calidad y cobertura sesiones 3g (int)

Bt_rsrq_rsrp	-periodo: fecha en formato MesAño (entero o int) -movil_e: número telefónico anonimizado (string o str) -median_rsrp: mediana de la potencia de la señal 4G recibida en el mes (float o flt) -p05_rsrp: percentil 5 de la potencia 4g recibida en el mes (flt) -p95_rsrp: percentil 95 de la potencia 4g recibida en el mes (flt) -median_rsrq: mediana de la calidad recibida en señal 4g en el mes (flt) -p05_rsrq: percentil 5 de la calidad 4g recibida en el mes (flt) -p95_rsrq: percentil 95 de la calidad 4g recibida en el mes (flt) -q_umbral_rsrp: cantidad de sesiones 4G en el mes bajo un umbral de potencia (int) -q_umbral_rsrq: cantidad de sesiones 4G en el mes bajo un umbral de calidad (int) -number_of_sessiones_4g: número de sesiones 4G en el mes (int) -percent_umbral_rsrp: porcentaje de sesiones 4G en el mes bajo un umbral de potencia (flt) - percent_umbral_rsrq: porcentaje de sesiones 4G en el mes bajo un umbral de calidad (int) -cob4buena_rate: porcentaje de sesiones 4G buenas sobre el total en el mes (flt) -cob4regularbuena_rate: porcentaje de sesiones 4G buenas y regulares sobre el total en el mes (flt) -cob4regular_rate: porcentaje de sesiones 4G regulares sobre el total en el mes (flt) -cob4regularmala_rate: porcentaje de sesiones 4G regulares y malas sobre el total en el mes (flt) -cob4mala_rate: porcentaje de sesiones 4G malas sobre el total en el mes (flt) -scorecobertura4g: score de 1-5 de la calidad y cobertura sesiones 4g (int)

Cdr_bt_datos_kpis_experiencia	-movil_e: número telefónico anonimizado (str) -n_celdas: número de celdas traficadas en el mes (int) -n_celdas_unicas: número de celdas únicas traficadas en el mes (int) -n_cambios_4g_3g: número de veces que pasa de tecnologia 4G a 3G en el mes -p_cambios_4g_3g: porcentaje de veces sobre el total de registros que pasa de 4G a 3G en el mes (flt) -trafico_subida: cantidad de datos cargados hacia la red en el mes (flt) -trafico_bajada: cantidad de datos descargados desde la red en el mes (flt) -trafico_total: suma de las 2 variables anteriores (flt) -mean_duracion: promedio de las duraciones de las sesiones en el mes (flt) -sum_duracion: suma total de todas las duraciones de las sesiones en el mes (flt) -p_celdas_4g: número de celdas 4G traficadas sobre el total de celdas traficadas en el mes (flt) -p_celdas_3g: análogo para celdas 3G -p_trafico_4g: porcentaje de trafico_total que ocurre en celdas 4G en el mes (flt) -p_trafico_3g: análogo para 3G -p_tiempo_4g: porcentaje del tiempo de las sesiones que ocurre en 4G en el mes (flt) -p_tiempo_3g: análogo para 3G (flt) -p_tiempo_4g_activo: análogo a p_tiempo_4g, pero con tráfico_total distinto a 0 (flt) -p_tiempo_3g_activo: análogo a p_tiempo_3g, pero con tráfico_total distinto a 0 (flt) -tiempo_activo: tiempo total de sesiones en donde trafico_total es distinto a 0 (flt) -tiempo_inactivo: análogo a tiempo_activo, pero con trafico_total igual a 0 (flt)
Cdr_bt_kpi_trafico_diario	-movil_e: número telefónico anonimizado (str) -dia: fecha en formato año-mes-dia (str) -clasificacion: tipo de tecnología de la celda (str) -cell_id_mas_traficada_dia_1: ID de la celda más traficada en el día (int) -cell_id_mas_traficada_dia_2: ID de la 2da celda más traficada en el día (int) -cell_id_mas_traficada_dia_3: ID de la 3ra celda más traficada en el día (int) -cell_id_mas_traficada_noche_1: ID de la

	celda más traficada en la noche (int) -cell_id_mas_traficada_noche_2: ID de la 2da celda más traficada en la noche (int) -cell_id_mas_traficada_noche_3: ID de la 3ra celda más traficada en la noche (int) -cantidad_celdas_distintas: cantidad de ID de celdas distintos traficados (juntando día y noche) (int) -cantidad_celdas_dia: análogo a cantidad_celdas_distintas, pero sólo en el día (str) - cantidad_celdas_noche: análogo a cantidad_celdas_distintas, pero sólo en la noche (str)
Bt_redes_celdas	-Cell_id: ID de la celda -cell_name: nombre de la celda (str) -zona: categoría de la zona en donde está la celda (str) -pop_m_id: nombre del POP (str) -comuna: comuna de la celda (str) -lat_wgs84: latitud de la celda (flt) -lon_wgs84: longitud de la celda (flt) -periodo: fecha en formato MesAño (int)
Redes_bt_calidad_3g_diario	-fecha: en formato año-mes-dia (str) -celda: nombre de la celda (str) -categoria: análogo a zona en bt_redes_celdas (str) -num_tti_util: factor numerador del PRB (o utilización) de la celda (flt) -den_ttt_util: factor denominador del PRB de la celda (flt) -num_nut_3_g_dl: factor numerador del NUT (o velocidad de descarga) de la celda (flt) -den_nut_3_g_dl: factor denominador del NUT de la celda (flt) -drop_cs: ratio de cortes en la señal de la celda (flt) -acc_cs: ratio de accesos en la señal de la celda (flt) -hs_users: ratio de cantidad de usuarios conectados a la celda (flt) -indisponibilidad: porcentaje de indisponibilidad de la celda (flt)

Redes_bt_calidad_4g_diario	-celda: nombre de la celda (str) -fecha: en formato año-mes-dia (str) -categoria: análogo a zona en bt_redes_celdas (str) -num_prb: factor numerador del PRB de la celda (flt) -den_prb: factor denominador del PRB de la celda (flt) -num_nut_4_g_dl: factor numerador del NUT de la celda (flt) -den_nut_4_g_dl: factor denominador del NUT de la celda (flt) -num_packet_discard_dl: factor numerador de del ratio de descarga de paquetes de la celda (flt) -den_packet_discard_dl: factor denominador de del ratio de descarga de paquetes de la celda (flt)
Bi_fct_clnt_pspg	-descr_movil_e: número telefónico anonimizado (str) -antigüedad_movil: antigüedad en meses del cliente (int) -fecha_proceso: en formato año-mes-dia (str)
Huawei_sdr_msisdn_ps	-movil_e: número telefónico anonimizado (str) -access_type: tipo de tecnología (str) -fecha_proceso: fecha en formato año-mes-dia (str) -chile_multimedia_responde_total_delay: delay total de la señal del cliente en la fecha (flt) -page_response_delay_ms_kpi: tasa de delay de respuesta de navegación (flt) -multimedia_responde_delay_ms_kpi: tasa de delay de contenido multimedia (flt) -page_response_success_rate_kpi: tasa de éxito en respuesta de navegación (flt) -webpage_download_speed_kbps_kpi: tasa de velocidad de descarga de contenido en la web (flt) -chile_page_sr_delay_msel: tasa de delay de navegación (flt) -video_streaming_xkb_start_delay_ms_kpi: tasa de delay en iniciar un video de streaming (flt) -file_sharing_response_delay: tasa de delay de respuesta en compartir archivos (flt) -client_dl_stream_tcp_packet_lossrate_kpi: tasa de pérdida de paquetes de stream (flt)

Anexo R. Vistazo Planilla Excel Caso de Negocio

Al mes 35 (seleccionado), y empezando en el mes 13, se observa un retorno (beneficio acumulado) positivo, esto es, 22 meses después reteniendo 30 clientes mensuales.

Supuestos Negocio	Valor	Comentario
Costo mejora promedio POP	70.000.000	Validado SmartCapex
Factura promedio	15.300	Validado Personas
% fuga promedio x POP	0,65%	Datos octubre 2023
WACC Real Mensual	0,84%	Dato negocio
Supuestos del proceso		
Delay liberación SmartCapex hasta mejora 12 meses		Validado SmartCapex
2023 se ejecutaron 500 proyectos mejora 4G		Validado SmartCapex
Supuestos del análisis		
Caso POP	AT295	
Fuga mensual	70	
Clientes mensual	7.000	
% Fuga	1,00%	
Potencial reducción fuga	30	
% fuga esperada post mejora	0,57%	

Periodo	Fugas retenidas	Total Fugas	Ingreso Mensua	Inversión	Beneficio acumulado
2025-01	30	30	415.163	70.000.000	(69.584.837)
2025-02	420	450	6.175.570	70.000.000	(133.409.267)
2025-03	450	900	12.248.255	70.000.000	(191.161.012)
2025-04	480	1380	18.624.214	70.000.000	(242.536.797)
2025-05	510	1890	25.294.602	70.000.000	(287.242.196)
2025-06	540	2430	32.250.725	70.000.000	(324.991.471)
2025-07	570	3000	39.484.043	70.000.000	(355.507.428)
2025-08	600	3600	46.986.168	70.000.000	(378.521.260)
2025-09	630	4230	54.748.857	70.000.000	(393.772.403)
2025-10	660	4890	62.764.014	70.000.000	(401.008.388)
2025-11	690	5580	71.023.687	70.000.000	(399.984.701)
2025-12	720	6300	79.520.066	70.000.000	(390.464.636)
2026-01	750	7020	87.869.965	70.000.000	(372.594.670)
2026-02	780	7380	91.606.622	70.000.000	(350.988.048)
2026-03	810	7740	95.274.928	70.000.000	(325.713.120)
2026-04	840	8100	98.875.764	70.000.000	(296.837.357)
2026-05	870	8460	102.409.998	70.000.000	(264.427.359)
2026-06	900	8820	105.878.491	70.000.000	(228.548.868)
2026-07	930	9180	109.282.092	70.000.000	(189.266.775)
2026-08	960	9540	112.621.643	70.000.000	(146.645.133)
2026-09	990	9900	115.897.973	70.000.000	(100.747.159)
2026-10	1020	10260	119.111.905	70.000.000	(51.635.254)
2026-11	1050	10620	122.264.250	70.000.000	628.996

Fuente: Desarrollado por el área de analytics de Entel

Anexo S. Ranking de correlación con el target para horizonte de tiempo

En este caso, el ranking muestra los 10 últimos y primeros registros en las figuras (a) y (b), respectivamente. El sufijo -1, -2 y -3 hace referencia a la variable agregada al mes previo a la fuga, a los 2 meses previos y a los 3, respectivamente. En general, en cada grupo de variable la variable agregada al mes previo correlaciona más en valor absoluto.

	Variable	Correlacion_churn=0
0	p_tiempo_4g-1	-0.151911
1	p_tiempo_4g_activo-1	-0.141072
2	p_tiempo_4g-2	-0.132073
3	p_celdas_4g-1	-0.126056
4	p_tiempo_4g_activo-2	-0.124479
5	periodo	-0.113189
6	p_celdas_4g-2	-0.111738
7	p_tiempo_4g-3	-0.100005
8	p_tiempo_4g_activo-3	-0.093543
9	p_trafico_4g-1	-0.091350
10	p_celdas_4g-3	-0.087995
11	p_trafico_4g-2	-0.084899
12	mean_duracion-2	-0.063644
13	p_trafico_4g-3	-0.063561
14	mean_duracion-1	-0.062503
15	mean_duracion-3	-0.057734
16	n_celdas_unicas-1	-0.054467
17	tiempo_inactivo-3	-0.051576
18	tiempo_inactivo-1	-0.049824
19	tiempo_inactivo-2	-0.046758

Anexo (a)

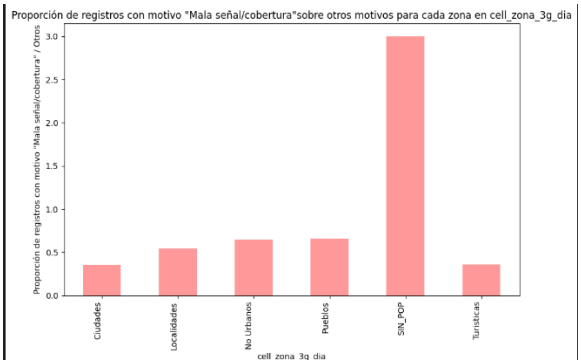
	Variable	Correlacion_churn=1
0	p_tiempo_3g-1	0.140138
1	p_cambios_4g_3g-1	0.132253
2	p_tiempo_3g_activo-1	0.131804
3	n_cambios_4g_3g-1	0.124905
4	p_tiempo_3g-2	0.121360
5	p_celdas_3g-1	0.118042
6	p_tiempo_3g_activo-2	0.117147
7	p_celdas_3g-2	0.103533
8	p_cambios_4g_3g-2	0.100937
9	n_cambios_4g_3g-3	0.097878
10	p_cambios_4g_3g-3	0.097812
11	p_tiempo_3g-3	0.093035
12	p_tiempo_3g_activo-3	0.090658
13	p_trafico_3g-1	0.089090
14	n_cambios_4g_3g-2	0.087061
15	p_celdas_3g-3	0.084715
16	p_trafico_3g-2	0.083879
17	p_trafico_3g-3	0.066120
18	n_celdas-1	0.033751
19	n_celdas-3	0.030390

Anexo (b)

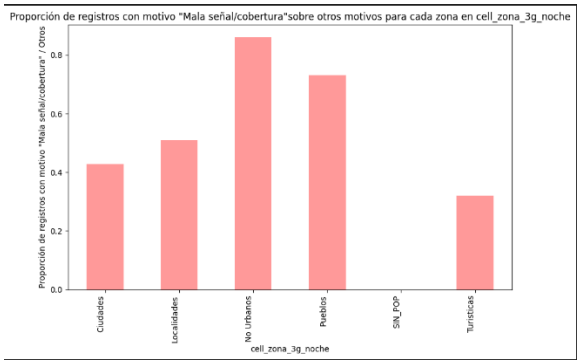
Anexo T. Vistazo del tablón inicial resultante desde el código

	periodo_po	movil_e	motivo	periodo_merge	tasa_dia_q_celdas_2g3g	tasa_noche_q_cel
1066	202305	6e6b64d96f1220ad720cec14b47d42943c8bc5c6e00feb...	Otros	202304	0.057897	
1067	202305	944256b76ebe4dd84b9d13ebb077687a1d0073b4f30b75...	Otros	202304	0.148936	
1068	202305	11dd4d8637c738ebe53a7f7bebfcf67f5073449316d98c...	Otros	202304	0.081081	
1069	202305	dc07f67240dc5cbdcf6c351339d74640e3aabd65e82226...	Mala señal/cobertura	202304	0.698276	
1070	202305	05a1ee29d4f82f66e95d9b94b38b73cfd4aa1f0bf0a609...	Otros	202304	0.125384	
...	...	...	...	...	...	
5556	202312	5eeccc6242e0990e9f102a6277732a9ad6931c700dfde2...	Otros	202311	0.172185	
5557	202312	3bb9c08b18d7954c4db2871103d4d1fc7fa7504e03fda2...	Otros	202311	0.048255	
5558	202312	10c902839e8425c224a9416b03e04ae738b2f39d34b3de...	Mala señal/cobertura	202311	0.235167	
5559	202312	da148599bf53bfb6960f9d7ad99152840942dbeac020f8...	Otros	202311	0.083744	
5560	202312	75a9927becc4faee527bf694f9cc23a39ae8dff067395f...	Mala señal/cobertura	202311	0.087273	

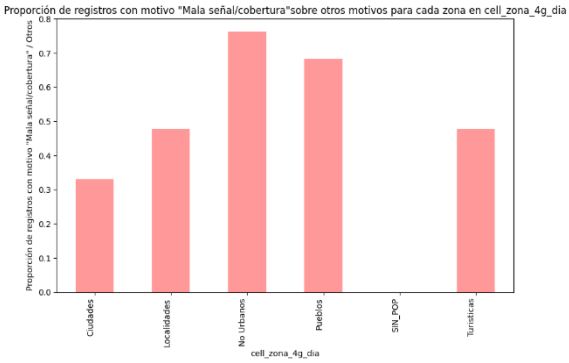
Anexo U. Proporción de fugados por red sobre el total, para las distintas categorías de celdas más usadas 3G y 4G en el día y en la noche



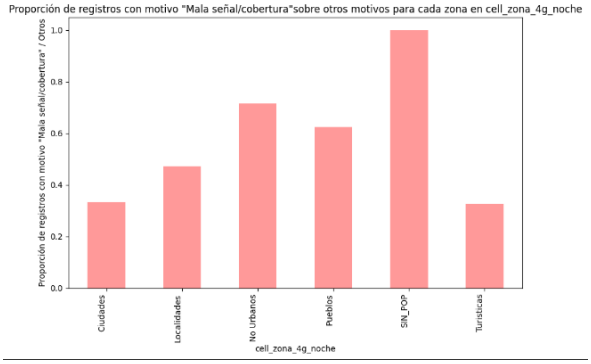
Anexo (a): Celda 3G día



Anexo (b): Celda 3G noche



Anexo (c): 4G día



Anexo (d): 4G noche

Anexo V. Variables eliminadas en la etapa de preparación de la data, junto con su procedencia

Variable	Tratamiento	Procedencia
cell_hs_users_3g_noche	Nulos	Atributo de celda más traficada 3G noche procedente de la tabla redes_bt_calidad_3g_diario
cell_avg_nut_dl_3g_noche	Nulos	Análogo a lo anterior
cell_avg_nut_dl_3g_dia	Nulos	Análogo a lo anterior, pero con la celda 3g día más usada
cell_avg_nut_dl_4g_noche	Nulos	Análogo a lo anterior, pero con la celda 4g noche y con la tabla redes_bt_calidad_4g_diario
sm_ps_4g_chile_multimedia_response_total_delay	Nulos	Tabla Huawei_sdr_msisdn_ps
sm_ps_3g_chile_multimedia_response_total_delay	Nulos	Tabla Huawei_sdr_msisdn_ps
sm_ps_4g_chile_page_sr_delay_msel	Nulos	Tabla Huawei_sdr_msisdn_ps
sm_ps_3g_chile_page_sr_delay_msel	Nulos	Tabla Huawei_sdr_msisdn_ps
Trafico_bajada	Correlaciones	Tabla cdr_bt_datos_kpis_experiencia
Score_cobertura_3G_4G_ponderado_tiempo_activo	Correlaciones	Creación propia
Score_cobertura_3G_4G_ponderado_p_tipo_celda	Correlaciones	Creación propia
P_tiempo_3g_activo	Correlaciones	Tabla cdr_bt_datos_kpis_experiencia
P_tiempo_4g	Correlaciones	Tabla cdr_bt_datos_kpis_experiencia
Score_cobertura_3G_4G_ponderado_trafico	Correlaciones	Creación propia
P_trafico_3g	Correlaciones	Cdr_bt_datos_kpis_experiencia
Cell_drop_cs_3g_noche	Desviación estándar	Atributo de celda más traficada 3G noche procedente de la tabla

		redes_bt_calidad_3g_diario
Cell_acc_css_3g_noche	Desviación estándar	Análogo a lo anterior
Cell_avg_tti_util_3g_noche	Desviación estándar	Análogo a lo anterior
Cell_zona_ciudad_4g_dia	Desviación estándar	Creación propia, en base a la categoría de la zona de la celda 4G día más usada, proveniente de la tabla redes_bt_calidad_4g_diario
Cell_zona_ciudad_4g_noche	Desviación estándar	Análogo a lo anterior, pero con la celda 4g noche más usada
Cell_zona_ciudad_3g_día	Desviación estándar	Análogo a lo anterior, pero con la celda 3g día y con la tabla redes_bt_calidad_3g_diario
Cell_zona_ciudad_3g_noche	Desviación estándar	Análogo a lo anterior, pero con la celda 3g noche

Anexo W. Hiperparámetros y parámetros de XGBoost y su descripción (obtenido de [29])

objective	Especifica la tarea de aprendizaje y el objetivo de aprendizaje correspondiente.
N_estimators	El número de rondas para ejecutar la capacitación (o entrenamiento).
learning_rate	Contracción del tamaño del paso utilizado en las actualizaciones para evitar el ajuste excesivo.
max_depth	Profundidad máxima de un árbol. El aumento de este valor hace que el modelo sea más complejo y que se sobreajuste con más probabilidad.
min_child_weight	Suma mínima de la ponderación de instancias (registros) necesaria en un elemento secundario. Conforme mayor sea el algoritmo, más conservador será.
subsample	La proporción de la submuestra de la instancia de capacitación. Esto evita el sobreajuste.
gamma	La reducción de pérdida mínima necesaria para realizar una partición mayor en un nodo de hoja del árbol. Conforme mayor sea, más conservador será el algoritmo.

colsample_bytree	Proporción de la submuestra de columnas cuando se construye cada árbol.
scale_pos_weight	Controla el equilibrio de las ponderaciones positivas y negativas. Resulta útil para las clases sin equilibrar.
nthread	Número de subprocesos paralelos utilizados para ejecutar *xgboost*.
seed	Semilla de número aleatorio.

Anexo X. Variables finales resultantes para modelar (más churn y periodo_po)

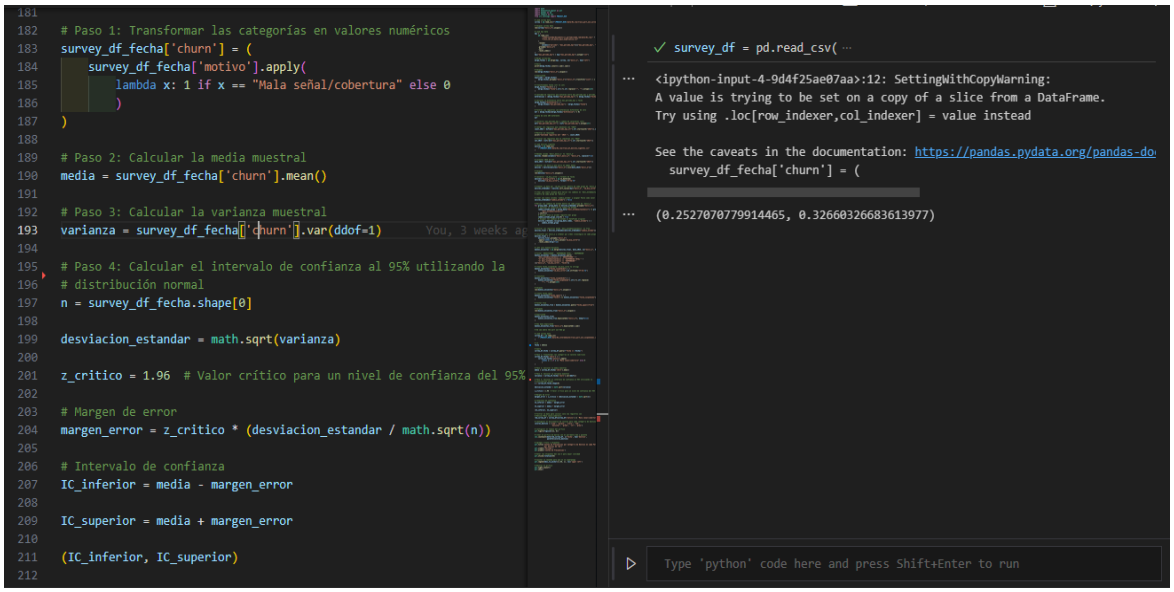
p_tiempo_4g_activo	tiempo_activo	tasa_dia_q_celdas_2g_3g: Tasa de uso en el día de celdas 2g/3g en el mes	tasa_noche_q_celdas_2g_3g: Tasa de uso en la noche de celdas 2g/3g en el mes	median_ecno
p05_ecno	p95_ecno	median_rscp	p05_rscp	cell_drop_cs_3g_dia: drop_cs de la celda 3G día más traficada
q_umbral_ecno	q_umbral_rscp	number_of_sessions_3g	percent_umbral_ecno	percent_umbral_rscp
cell_hs_users_3g_dia: hs_users de la celda 3G día más usada.	cob4regularbuena_rate_3g	cell_avg_tti_util_3g_dia: Promedio de la utilización mensual de la celda 3G día más usada.	cob4regularmala_rate_3g	cell_avg_nut_dl_4g_dia: Promedio mensual de NUT de la celda 4g más usada en el día.
scorecobertura3g	median_rsrp	p05_rsrp	p95_rsrp	cell_avg_packet_discard_dl_4g_dia: packet_discard_dl de la celda 4G día más usada.

Antigüedad_movil	p95_rsrq	q_umbral_rsrp	q_umbral_rsrq	sm_ps_3g_page_response_delay_ms_kpi
percent_umbral_rsrp	percent_umbral_rsrq	sm_ps_3g_multimedia_response_delay_kpi	sm_ps_4g_page_response_delay_ms_kpi	sm_ps_4g_multimedia_response_delay_kpi
cob4regularmala_rate_4g	cob4mala_rate_4g	scorecobertura4g	n_celdas	sm_ps_4g_page_response_success_rate_kpi
n_cambios_4g_3g	sm_ps_4g_webpage_download_speed_kbps_kpi	sm_ps_4g_video_streaming_xkb_start_delay_ms_kpi	churn: Variable binaria que es 1 si la fuga fue por red y 0 si no.	trafico_total
sm_ps_4g_client_dl_stream_tcp_packet_lossrate_kpi	cell_acc_css_3g_dia: acc_cs de la celda 3G día más usadas.	p_celdas_4g	p_celdas_3g	Periodo_po: Mes del porto ut del cliente
p_tiempo_3g	cell_avg_packet_discard_dl_4g_noche: Análogo a cell_avg_packet_discard_dl_4g_dia, pero de noche.			

Anexo Y. Hiperparámetros finales modelo XGBoost

```
params = {
    'objective': 'binary:logistic',
    'learning_rate': 0.05662929606402181,
    'max_depth': 3,
    'min_child_weight': 2,
    'subsample': 0.6189844918742183,
    'gamma': 0.2788995088939747,
    'colsample_bytree': 0.7559383386201027,
    'scale_pos_weight': 1.4444836389457774,
    'nthread': 4,
    'seed': 42
}
```

Anexo Z. Código para obtener el intervalo de confianza, en este caso, de diciembre



```
# Paso 1: Transformar las categorías en valores numéricos
survey_df_fecha['churn'] = (
    survey_df_fecha['motivo'].apply(
        lambda x: 1 if x == "Mala señal/cobertura" else 0
        )
)

# Paso 2: Calcular la media muestral
media = survey_df_fecha['churn'].mean()

# Paso 3: Calcular la varianza muestral
varianza = survey_df_fecha['churn'].var(ddof=1)

# Paso 4: Calcular el intervalo de confianza al 95% utilizando la
# distribución normal
n = survey_df_fecha.shape[0]

desviacion_estandar = math.sqrt(varianza)

z_critico = 1.96  # Valor crítico para un nivel de confianza del 95%

# Margen de error
margen_error = z_critico * (desviacion_estandar / math.sqrt(n))

# Intervalo de confianza
IC_inferior = media - margen_error

IC_superior = media + margen_error

(IC_inferior, IC_superior)
```

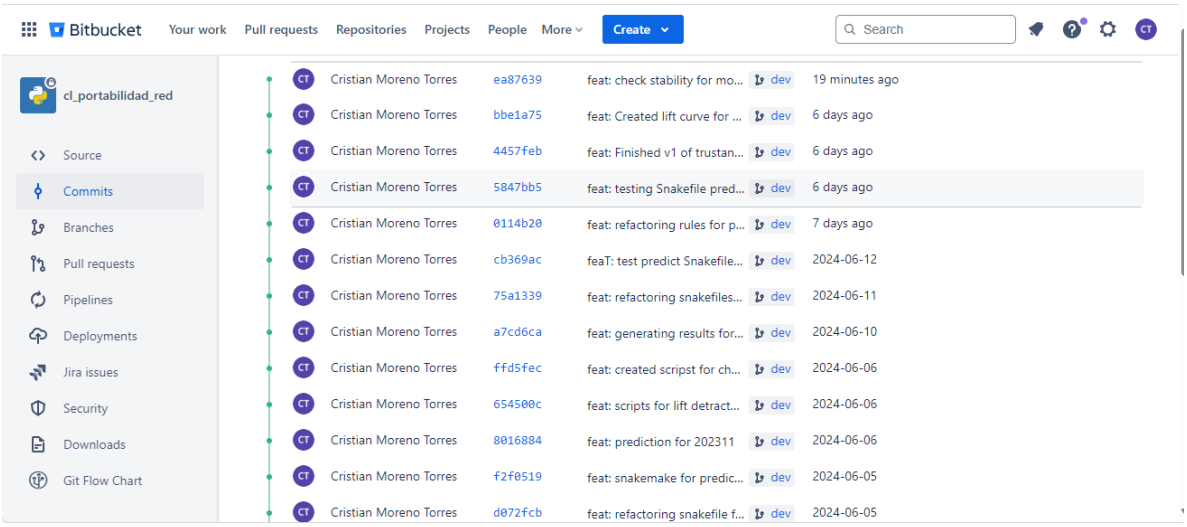
```
survey_df = pd.read_csv(...
<ipython-input-4-9d4f25ae07aa>:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-do
  survey_df_fecha['churn'] = (

(0.2527070779914465, 0.3266032668361397)

Type 'python' code here and press Shift+Enter to run
```

Anexo A1. Vistazo del respaldo del desarrollo en el repositorio virtual en Bitbucket



Anexo B1. Vistazo de comentarios metódicos dentro de los scripts

```
# Load master table
df_master_table = pd.read_csv(
    f"{PROJECT_DIR}/data/03_processed/{stage}/{periodo}/"
    + "master_table_v1_clean.csv"
)

# Drop columns NAs
df_master_table = df_master_table.drop(
    columns=[
        "cell_hs_users_3g_noche",
        "cell_avg_nut_dl_3g_noche",
        "cell_avg_nut_dl_3g_dia",
        "cell_avg_nut_dl_4g_noche",
        "sm_ps_4g_chile_multimedia_response_total_delay",
        "sm_ps_3g_chile_multimedia_response_total_delay",
        "sm_ps_4g_chile_page_sr_delay_msel",
        "sm_ps_3g_chile_page_sr_delay_msel",
    ]
)

# Separate numeric and non-numeric columns
numeric_cols = df_master_table.select_dtypes(include="number").columns
non_numeric_cols = df_master_table.select_dtypes(exclude="number").columns

# Handle missing values for numeric columns
numeric_imputer = SimpleImputer(strategy="mean")
df_master_table[numeric_cols] = numeric_imputer.fit_transform(
    df_master_table[numeric_cols]
)
```

Anexo C1. Vistazo del documento Readme del proyecto