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

MODELO PREDICTIVO SOBRE LA PERMANENCIA DE LAS NUEVAS LÍNEAS MÓVILES POSTPAGO DEL SEGMENTO DE PERSONAS EN ENTEL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

FRANCISCA BELÉN SANHUEZA CONTRERAS

PROFESORA GUÍA:
LORETO MARTÍNEZ GIMÉNEZ

PROFESOR CO-GUÍA:
ALEJANDRO MUÑOZ ROJAS

COMISIÓN:
BLAS DUARTE ALLEUY

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: Ingeniera Civil Industrial
POR: Francisca Belén Sanhueza Contreras
FECHA: 2023
PROFESORA GUÍA: Loreto Martínez Giménez

El sector de las telecomunicaciones en Chile destaca por la alta competencia que existe entre las 4 principales empresas Entel, Movistar, Claro y WOM. Esto fue intensificado a partir de la Ley de Portabilidad, que facilita la migración de clientes de una compañía a otra. A partir de esto, nace la necesidad de reducir la tasa de fuga de los clientes, apuntando principalmente a una mejor retención, lo cual es especialmente relevante ya que retener un cliente es más rentable que captar un nuevo cliente.

Considerando que, actualmente la empresa cuenta con un modelo que entrega información de la fuga a nivel general, esto es, el porcentaje de fuga de clientes del mes siguiente, nace la oportunidad de realizar predicciones personalizadas sobre el tiempo de permanencia total de un cliente en la empresa. Este último enfoque permitiría generar campañas de retención más eficientes, mediante estrategias preventivas y personalizadas. A partir de lo anterior, se implementó un *modelo con el objetivo de predecir la permanencia de las líneas móviles pospago del segmento de personas al momento de la habilitación*.

Para el desarrollo del proyecto se aplicó una metodología CRISP-DM ligeramente modificada. Esta consta de 5 etapas: entendimiento del negocio, entendimiento de los datos, procesamiento de datos, modelamiento y traspaso del modelo. Cabe destacar que esta metodología ha mostrado buenos resultados en proyectos anteriores dentro de la empresa Entel.

El proyecto se llevó a cabo mediante el modelamiento de datos a través de un Análisis de Supervivencia, el que entrega información respecto a la permanencia de las nuevas líneas móviles. Más precisamente, la esperanza de 'vida' de cada línea móvil, la probabilidad de fuga en meses críticos y los coeficientes asociados a las variables predictivas. Además, se corroboró que al separar las muestras en deciles dados por las esperanzas de vida de cada cliente, el orden inducido se mantiene a través del tiempo. En resumen, se puede afirmar que se cumplió con los objetivos específicos y general del proyecto.

Por último, resulta relevante mencionar que, durante el desarrollo del proyecto, se constataron nuevos casos de uso como, por ejemplo, la identificación de variables asociadas a mejores ventas, esto es, ventas de líneas con mayor supervivencia a través del tiempo.

Tabla de Contenido

1. ANTECEDENTES GENERALES DE LA EMPRESA

Entel S.A. (Empresa Nacional de Telecomunicaciones S.A.) es una empresa de telecomunicaciones que fue fundada en 1964 como una empresa estatal y que fue privatizada en el año 1992. La empresa nace con el objetivo de entregar servicios de telefonía de larga distancia nacional e internacional y servicios telegráficos. Actualmente ofrece servicios de conectividad móvil y fija, así como una amplia gama de servicios TI y digitales.[1]

Entel cuenta con presencia internacional en Perú desde el año 2004. Según lo publicado por la empresa, entre ambos países suman alrededor de 20,4 millones de abonados móviles y 3.200.000 millones de pesos en ingresos consolidados[2].

1.1 PROPÓSITO Y VALORES

Como se declara en la Memoria Anual de Entel 2022, el propósito de la empresa es 'Poder acercar las infinitas posibilidades que da la tecnología y así transformar responsablemente la sociedad.' En línea con lo anterior, se describe la misión de la empresa como 'Conectar a las personas y acercar la tecnología, habilitar su uso, y acercar las infinitas posibilidades que brinda la tecnología a la vida cotidiana de las personas.' [3]

Lo anterior se sostiene bajo una cultura empresarial que se basa en 6 valores y pilares fundamentales: calidad, orientación al cliente, responsabilidad, trabajo en equipo, valorar la diversidad, ser eficientes y atreverse a probar nuevas formas de hacer las cosas e ir por más siempre.

1.2 ORGANIGRAMA

A continuación se presenta un diagrama que representa la organización de las principales áreas de la empresa, siendo CVM, equipo de Customer Value Management en la cual se desarrolla el Trabajo de Título.

---

[1] Información publicada en el sitio web de la empresa, www.entel.cl, [consulta: 20 Octubre 2023]
[2] Información publicada en la sección de información corporativa del sitio web de la empresa,https://informacioncorporativa.entel.cl/ [consulta: 22 Octubre 2023]
[3] Según lo declarado en la Memoria Integrada 2022 de la empresa.

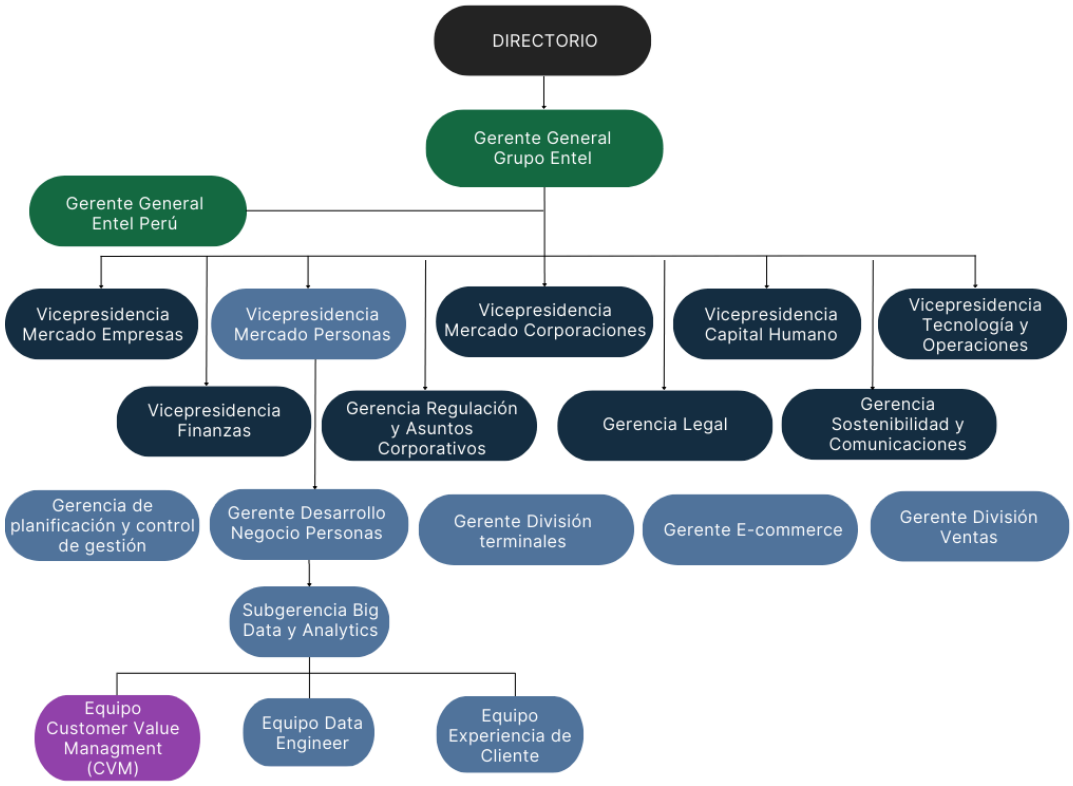

Figura 1: Organigrama Entel, elaboración propia[4]

En la Figura 1 se observa el organigrama de la organización. La compañía es encabezada por el Gerente General. Tiene bajo su liderazgo todas las Vicepresidencias y Gerencias de la empresa, las que se distribuyen según los segmentos de clientes: mercado personas, mercado empresas, mercado corporaciones.

La Gerencia de Desarrollo de Negocios, tiene por objetivo velar por el cumplimiento de las distintas dimensiones del negocio, tanto móvil como hogar. Su rol es velar por una mejor experiencia para el cliente y definir objetivos de medio y largo plazo. A su cargo están distintos equipos que buscan crear estrategias de negocios, iniciativas y desarrollos según el tipo de mercado que abordan.

La Subgerencia de Big Data y Analytics, tiene por objetivo apoyar la toma de decisiones mediante la explotación avanzada de grandes volúmenes de datos. Dentro de esta área conviven tres equipos. El equipo de experiencia de clientes, equipo de Data Engineers, y por último, el equipo de Customer Value Management o CVM, en el

[4] Organigrama realizado a partir de la información obtenida en la Memoria Anual 2022 publicada por la empresa.

que se enmarca el Trabajo de Título. Este último equipo busca crear y gestionar campañas personalizadas, entregando la mejor oferta para cada cliente del mercado de personas en el momento preciso. Además, desarrollan herramientas que permiten caracterizar y predecir comportamientos de los clientes para focalizar los recursos de la forma más eficiente.

1.3 SECTOR INDUSTRIAL

La actividad económica de las comunicaciones y servicios de información abarca la producción, distribución y transmisión de información y productos culturales, así como tecnología de la información y procesamiento de datos. Para el año 2022, la industria representó el 6,8% del PIB Chileno.[5]

La industria en la cual se enmarca el Trabajo de Título corresponde a la industria de las telecomunicaciones, esta incluye servicios de telefonía fija, de larga distancia, televisión de pago, de Internet, servicios intermedios de telecomunicaciones, transmisión de datos y radiocomunicaciones móviles. Según el Ministerio de Telecomunicaciones y Transportes, la actividad predominante en esta industria es la telefonía móvil. El Trabajo de Título se realiza para la telefonía móvil, por ello, los antecedentes que se presentan a continuación se centran en este segmento.

El último informe de la Subtel indica que, a diciembre de 2022, el sector alcanzó los 58,8 millones de abonados en los distintos servicios de telecomunicaciones, de los cuales 48,8 millones corresponden a abonados en servicios de telefonía e internet móvil.

Las 4 empresas que acaparan el mercado en los servicios móviles son Entel, Movistar, Wom y Claro. Estas concentraron el 96% de la participación de mercado chileno en número de abonados respecto a telefonía móvil [31]. A continuación, se presenta un gráfico que muestra la distribución de la participación de mercado del segmento móvil postpago para fines del año 2022.

[5] A partir de la información revelada por el portal de estadística, Statista.

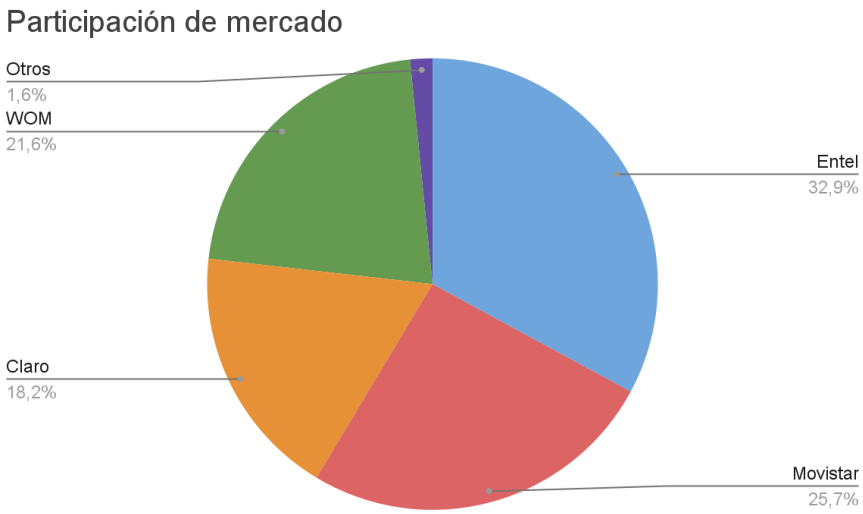

Figura 2: Participación de mercado en el segmento móvil pospago para el año 2022.
Fuente: SUBTEL

1.4 ANÁLISIS PESTEL

A continuación se presenta un análisis PESTEL, en el cual se analizan los factores externos que influyen en los servicios móviles de la industria de telecomunicaciones.

1.4.1 POLÍTICO

El gobierno de Chile estableció una estrategia de transformación digital para 2035 en la que se definieron metas en infraestructura digital, habilidades y derechos digitales y ciberseguridad. Específicamente, en cuanto a servicios de telecomunicación, la estrategia propone alcanzar 90% de cobertura 5G en 2025, 98% en 2030 y 100% en 2035. Para lo cual se estiman inversiones de infraestructura 5G por US$6.000 millones para el período 2022-2031. Por otro lado, la estrategia prevé una penetración móvil de 80% para 2025 y de 98% para 2035 [34]. Estos antecedentes permiten evidenciar la disposición estatal a fomentar y facilitar el despliegue de tecnologías móviles a través de un plan de modernización digital.

1.4.2 ECONÓMICO

Después de la pandemia las tasas de inflación conocieron un fuerte crecimiento en el mundo y particularmente en Chile. En particular en Chile la inflación anual alcanzó un 12,8% en 2022 según el Diario Financiero. A pesar de que el panorama general ha mejorado desde 2022 según el informe del Banco Mundial, este mismo destaca que “En Chile, el PIB real se contrajo un 1% interanual en el primer semestre de 2023 y la tasa de desempleo aumentó 0,7 puntos porcentuales interanuales en junio de 2023”.

1.4.3 SOCIAL

En Chile, se identifican brechas digitales que afectan la demanda de servicios de telecomunicaciones y obstaculizan el desarrollo digital del país. La OCDE señala que estas brechas se manifiestan principalmente en diferencias significativas en el acceso según el nivel educativo, la edad, el ingreso del hogar, y la ubicación geográfica.

Con el objetivo de reducir esta brecha el gobierno lanzó el programa Brecha Digital Cer0 2022- 2025, cuyo objetivo central es que todos los habitantes del país tengan acceso a conectividad, independiente del lugar del país en que viven o de las posibilidades económicas que tengan. [18]

1.4.4 TECNOLÓGICO

La última tecnología en servicios móviles, corresponde a la conexión móvil 5G, lanzada en enero de 2022, experimentó un notable crecimiento del 560% en marzo, alcanzando 2,8 millones de dispositivos conectados en septiembre. Este aumento contrasta con la caída de las conexiones 4G, que pasaron de 20,3 millones en la primera mitad del año pasado a 18,4 millones en la primera mitad de este año.[6]

El ministro de Transportes y Telecomunicaciones, Juan Carlos Muñoz declaró: “El nivel de conexiones a la red 5G en los primeros cuatro meses del año demuestra que los chilenos quieren más y mejor tecnología. Pero también nos desafía a llegar con ella a to*dos los rincones del país, a acortar las brechas con las grandes ciudades.*

[6] Según un análisis publicado por el Ministerio de Transporte y Telecomunicaciones. *Especial Análisis Nueva Tecnología 5G en Internet Móvil y crecimiento Tecnología Fibra en Internet Fija. 2022.*

1.4.5 ECOLÓGICO

El sector de las telecomunicaciones representó el 2.6% de todas las emisiones globales de dióxido de carbono (CO2) en el año 2020. Desde una perspectiva económica, el gasto energético constituye entre el 20% y el 40% de los costos totales de un proveedor de servicios de telecomunicaciones. Aunque estos porcentajes pueden fluctuar en función de las circunstancias específicas, se observa que el 73% de la energía utilizada por los operadores se destina a las radiobases, mientras que el 9% se consume en los centros de datos.[7]

1.4.6 LEGAL

A continuación se presentan las principales legislaciones que rigen la industria. La ley General de Telecomunicaciones y La Ley de Portabilidad Numérica.

La Ley General de Telecomunicaciones (Ley N°18.168) establece un régimen libre e igualitario de acceso a las telecomunicaciones. La Subsecretaría de Telecomunicaciones es la entidad encargada de la regulación y supervisión de las empresas del sector, además de la propuesta de políticas nacionales y el control de su implementación. Por otro lado, es relevante mencionar que la definición de las tarifas de los productos o servicios ofrecidos por la empresa son definidos por ésta, salvo calificación expresa del Tribunal de Defensa de la Libre Competencia, en aquellos casos expresamente contemplados por la ley [4].

Por otro lado, los servicios de telecomunicaciones se rigen bajo la Ley N°20.471, Ley de Portabilidad numérica, implementada en 2011, que permite a usuarios de servicios de telefonía cambiarse de compañía manteniendo el número. Esta ley permite a los consumidores poder decidir entre las distintas compañías sin restricciones lo cual aumenta la competitividad en la industria [3]. Según el último *Repo*rte de Portabilidad Mensual de Subtel, desde que comenzó a regir esta ley en diciembre de 2011 hasta julio del 2023 se han portado más de 33 millones de números móviles (33.844.393), de los cuales el 95,6% (32.346.011) corresponden a portaciones de números móviles.

[7] Información obtenida a partir de publicación del diario El Economista, https://www.eleconomista.com.mx/opinion/Telecomunicaciones-y-medio-ambiente-20211203-0026.html. [Consulta: 19 Noviembre 2023]

A partir de estos antecedentes, se evidencia que, desde la implementación de la ley, existe un mayor poder de negociación por parte de los clientes en esta industria, debido a que se eliminan las restricciones para cambiarse libremente entre las compañías.

1.5 OPORTUNIDAD DE MEJORA

En la Memoria Anual de Entel se especifica la importancia de "*ser líderes indiscutidos en experiencia de clientes en cada categoría en la que participamos, por medio de servicios diferenciadores, innovaciones tecnológicas y una* ***atención que se ajusta a las necesidades de cada cliente****, generando una ventaja competitiva y sostenible que nos permita seguir expandiendo el negocio en un mercado cada vez más desafiante. Esto requiere* ***identificar en forma certera las necesidades de nuestros usuarios y trabajar entre todas las áreas para satisfacerlas*** *con un servicio superior acorde con los atributos de nuestra experiencia: simple, cercana y sorprendente*".

Según las declaraciones anteriores, la experiencia y fidelización de los clientes son aspectos relevantes para la empresa. Además, para poder brindar un servicio que se adapte a las necesidades de los clientes, es fundamental entender cuáles son esas necesidades a través de evidencia empírica.

El diccionario de Oxford define la personalización como "La acción de diseñar o producir algo que cumple con los requisitos individuales de alguien". En la práctica, la personalización se considera un proceso diseñado para crear una interacción relevante y personalizada con el fin de mejorar la experiencia del cliente [8]. Específicamente, en el proceso se utiliza información basada en los datos personales y de comportamiento de cada cliente para ofrecer una experiencia superior.

Una forma de abordar la fidelización y la retención es ofrecer productos y servicios según la permanencia de las líneas móviles. En el contexto de este Trabajo de Título, se define la permanencia como la cantidad de tiempo (en meses) que una línea mantiene un contrato de suscripción móvil con la empresa, esto es, suscripción postpago.

[8] Shobhana Chandra, Sanjeev Verma, Weng Marc Lim, Satish Kumar, Naveen Donthu 2022. Personalization in personalized marketing: Trends and ways forward

Al determinar la permanencia es posible definir la mejor estrategia para cada cliente. Por ejemplo, para un cliente con una estimación de una alta permanencia en la empresa, sería conveniente ofrecer beneficios que aumenten su fidelización. Por el contrario, para un cliente con una estimación de permanencia baja, lo más eficiente sería desplegar una campaña de retención.

Para caracterizar el tipo de cliente se puede predecir la permanencia de cada cliente, es decir, en cuánto tiempo se fugará, ya que esta información permite tener una estimación del tipo de relación con la empresa por ejemplo de largo, mediano o corto plazo. Esto le permitirá al área de marketing, tomar decisiones de fidelización o retención, ya sea para controlar la fuga o aumentar la satisfacción.

Por otro lado, se espera que la información entregada permita generar estrategias para la adquisición de líneas de alta calidad, es decir, aquellas líneas con una alta permanencia en la empresa, para esto se espera tener información respecto a cuáles son las variables que afectan, ya sea positiva o negativamente en la tasa de riesgo de fuga, y de esta forma, potenciar aquellos factores que influyen en una mejor venta y controlar aquellos que están asociados a ventas de líneas de peor calidad.

Actualmente, la empresa cuenta con un modelo predictivo que entrega el porcentaje de la base de clientes que se fugará el próximo mes. Sin embargo, el modelo no especifica cuáles son esos clientes, no entrega mayor información respecto a meses posteriores y tampoco explicita cuáles son las variables influyen en dicho porcentaje. A partir de esto nace la oportunidad de diseñar un modelo que entregue mayor información y permita llevar a los equipos correspondientes llevar a cabo estrategias personalizadas.

Debido a la forma en la que se realizan las predicciones actualmente, las estrategias de retención se despliegan para toda la base de clientes cuando aumenta el porcentaje de fuga estimado para el mes siguiente. Hay dos elementos que se pueden optimizar en este caso. En primer lugar, lo más eficiente sería desplegar las campañas de retención sólo para aquellos clientes que presentan riesgo de fuga. Por otro lado, aumentar el márgen de tiempo con el que se despliega la campaña, esto permitiría implementar estrategias con enfoque preventivo, más que reactivo.

Además, es importante mencionar que, a lo largo del desarrollo del proyecto, se definió un nuevo caso de uso para la información proporcionada por el modelo, relacionado con la detección de líneas fraudulentas. Este fenómeno se refiere a líneas móviles que son habilitadas con el único objetivo de que el vendedor obtenga el beneficio asociado a la comisión por venta. Estas líneas luego son deshabilitadas al primer mes por no pago. Es directo que, para esto se deben identificar dichas líneas que presentan la menor permanencia dentro de las predicciones.

2. JUSTIFICACIÓN DEL PROYECTO

Para abarcar las oportunidades planteadas en la sección anterior, se decidió la realización de una herramienta analítica que permita entregar información respecto a las necesidades de los clientes en función de su tiempo de vida en la compañía.

El modelo de predicción proporcionará información a nivel individual para cada línea móvil registrada en la base de datos, identificando aquellas con riesgo de abandonar el servicio en meses posteriores, de esta manera se podrá discriminar aquellas líneas y clientes con mayor probabilidad de fuga. Esto permitirá asignar los recursos de retención y fidelización eficientemente.

Sumado a lo anterior, este modelo se implementará en el segmento móvil postpago del Mercado de Personas. Este segmento es especialmente relevante dentro de la empresa debido al volumen de usuarios y al aporte en ingresos. En efecto, el mercado de personas representa aproximadamente el 98.25% del total de clientes en todos los segmentos. Además, los usuarios han mostrado una preferencia a los planes postpago, donde el cociente entre la cantidad de usuarios postpago sobre la cantidad de usuarios prepago ha mostrado un aumento, demostrando una variación positiva de 5.6% el año 2022 respecto al anterior.[9]

Por otro lado, Entel corresponde a una de las empresas con menor número de portaciones netas históricas en el segmento de telefonía móvil (ver Figura 3 en Anexos), esto es, el número de usuarios recibidos, menos los usuarios portados desde la compañía. Específicamente en el segmento móvil postpago, la empresa ha presentado un balance positivo los últimos tres años, sin embargo, esta no fue siempre

[9] Información sobre los ingresos de Clientes móviles postpago y prepago en 2021 y 2022 a partir de lo publicado en la Memoria Anual Entel 2022

la tendencia, específicamente entre 2016 y 2019, tuvo un balance negativo. Más aún, corresponde a una de las 6 empresas con mayor variación en las portaciones netas a través de los años. (ver Figura 4 en Anexos).

Cabe mencionar que la estrategia de retención es clave, dado que captar nuevos usuarios es considerablemente más costoso, en efecto, se estima que, con un aumento en las tasas de retención de clientes de tan solo un 5%, el valor neto presente promedio de un cliente aumenta en un 35%[10] Esto último es especialmente relevante en un mercado donde los clientes tienen un alto poder de elección a partir de la ley de portabilidad.

3. ROL DE LA ESTUDIANTE

3.1 DESCRIPCIÓN DEL ÁREA DE TRABAJO

El Trabajo de Título se realiza en la subgerencia Big Data y Analytics, en el que se llevan a cabo tareas de analítica de datos, lo que permite a otras áreas de la empresa tomar decisiones en base a sus resultados.

El equipo de Analytics, en el que se realiza el proyecto, lleva a cabo tareas de analítica de datos para el equipo de Gestión del valor del cliente, el cual a partir de ahora se denominará CVM por sus siglas en inglés. Éste determina la forma de distribución de las campañas de marketing. Por lo tanto, el análisis de datos llevado a cabo por este equipo dentro de Analytics, debe entregar información relevante que permita a CVM optimizar la gestión de dichas campañas.

3.2 ORGANIZACIÓN DEL TRABAJO

El desarrollo y ejecución de este proyecto recae primordialmente en la responsabilidad de la estudiante, quien actúa como la figura central encargada de llevar a cabo las tareas y actividades necesarias para la realización del proyecto. A pesar de la autonomía conferida, la estudiante cuenta con el respaldo constante de la tutora en la empresa. Esta figura desempeña un papel crucial al proporcionar orientación estratégica y entregar retroalimentación en reuniones programadas semanalmente.

[10] Reichheld and Teal, Harvard Business School Press, 1996. The Loyalty Effect: The Hidden Force behind Growth, Profits, and Lasting Value.

4. OBJETIVO GENERAL

El objetivo general del proyecto es diseñar un modelo de predicción de permanencia de las nuevas líneas móviles postpago en Entel, con el fin de entregar información que permita al equipo de Customer Value Management generar estrategias personalizadas de retención y fidelización de clientes.

5. OBJETIVOS ESPECÍFICOS

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

a. Entender las necesidades del equipo de CVM para establecer las bases del proyecto.
b. Reunir, preparar y estructurar los datos de los clientes postpago para el desarrollo del modelo.
c. Determinar la configuración óptima del modelo que permita obtener el resultado más robusto.
d. Documentar el modelo y realizar una capacitación a un integrante del equipo para hacer efectivo el traspaso del conocimiento generado a la empresa.

6. ALCANCES

A partir de los objetivos anteriores, se definen los límites de lo que se abordará en el proyecto, estableciendo aquellos aspectos que no pertenecen al alcance de este trabajo.

En primer lugar, es importante especificar que el modelo estima la permanencia de **líneas móviles** y no de clientes. Un cliente puede estar asociado a más de una línea móvil. Para ello, el modelo predice la permanencia **al momento de la habilitación** de la línea, de forma que, sólo se consideran datos asociados a este periodo (por ejemplo: edad del titular, monto asociado a la línea, género del titular, etc.) y no se incorporan datos relativos al comportamiento en meses posteriores (Por ejemplo: cantidad de datos traficados cada mes, número de reclamos, etc.). La realización de un modelo que integre dichas fluctuaciones quedó descartado pues

requiere el procesamiento de una cantidad de datos significativamente superior y el tiempo disponible para la realización del proyecto impedía llevarlo a cabo.

Por otro lado, sólo se consideran variables provenientes de información presente en las bases de datos de la empresa. Así, el análisis no incorpora información respecto a factores externos, como son, por ejemplo, los precios y descuentos que ofrecen los competidores en el periodo de estudio. Tampoco se considera información sociodemográfica adicional a la almacenada en las bases de la empresa. Si bien esta información puede afectar la permanencia del cliente y ser significativa para el modelo, considerando que los datos no son de fácil acceso y las limitaciones de tiempo, no se integraron.

Además, tal como se comentó en secciones anteriores, dado que el trabajo se realiza para el equipo de Analytics dentro de la Vicepresidencia de Personas, el estudio se concentra en el segmento más grande de esa unidad, es decir, líneas móviles postpago. Sin embargo, este análisis podría ser fácilmente replicable en otras unidades de servicio postpago, ya sea telefonía fija o para empresas. El único caso en el que el modelo necesitaría grandes modificaciones sería para analizar a los clientes de telefonía prepago.

Finalmente, es importante precisar que el proyecto sólo considerará el traspaso de la información al equipo de Analytics, no así la implementación del modelo. Esto se debe principalmente a la limitaciones de tiempo para el Trabajo de Título y de acceso a plataformas, restringidas sólo para trabajadores de Entel.

7. MARCO CONCEPTUAL

Es necesario precisar el significado de *fuga* en este contexto. Este término generalmente posee dos interpretaciones: la primera sería que el cliente deja de interactuar con la compañía en un intervalo de tiempo previamente determinado y la segunda sería la cancelación del contrato que mantenía el cliente con la empresa[11].

[11] Chandar and Krishna, 2006. Modeling churn behavior of bank customers using predictive data mining techniques.

En el contexto de este proyecto, la definición se aproxima más a la segunda acepción, con un pequeño matiz que, en vez de realizar el análisis con respecto a los clientes, este se realiza con respecto a las líneas móviles postpago de cada cliente. Más precisamente, el concepto *fuga* dice relación con el fin de un contrato de suscripción asociado a una línea, sin importar motivo de esta cancelación: portabilidad, migración postpago a prepago, deshabilitación definitiva de la línea móvil, entre otros. Sin embargo, por simplicidad, se hará referencia de ambas formas a lo largo del informe: fuga de línea móvil y fuga de cliente.

7.1 DISCIPLINA DEL PROYECTO

El desarrollo del proyecto se lleva a cabo mediante el ejercicio de la disciplina de la ciencia de datos, que corresponde a un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos, en este caso particular, de una empresa de telecomunicaciones[12].

Existen distintos enfoques analíticos que el cientista de datos puede utilizar para extraer información, estos son: análisis predictivo, análisis descriptivo y análisis prescriptivo. El análisis predictivo permite hacer pronósticos fundamentados en datos históricos y cuyo objetivo es, principalmente, determinar las probabilidades de que algo suceda en el futuro. A su vez, el análisis descriptivo permite reconocer tendencias y relaciones en datos históricos. Finalmente, el análisis prescriptivo corresponde a la utilización de datos para determinar un curso de acción óptimo considerando todos los factores relevantes obtenidos en el análisis descriptivo y predictivo. El Trabajo de Título se enmarca en el contexto de la analítica descriptiva y predictiva, con especial énfasis en esta última.

7.2 METODOLOGÍAS

Existen distintas metodologías utilizadas en el contexto de desarrollo de proyectos en el campo de la ciencia de datos, entre ellas, CRISP-DM, SEMMA y KDD. Cada una de ellas aborda de manera distinta aspectos clave como lo son: la flexibilidad, iteraciones, enfoque en el negocio y herramientas para su desarrollo. A

[12]Amazon. Ciencia de Datos. https://aws.amazon.com/es/what-is/data-science/#:~:text=La%20ciencia%20de%20datos%20es,analizar%20grandes%20cantidades%20de%20datos [Consulta: 21 Noviembre 2023]

continuación, se muestra una tabla que resume y compara las tres metodologías mencionadas anteriormente (Tabla completa en Anexos).

Aspecto	CRISP-DM	SEMMA	KDD
Flexibilidad	**Alta** **Estructura Modular** Cada fase se considera como un módulo independiente con objetivos claros, lo que permite adaptar el enfoque según las necesidades específicas del proyecto	**Moderada** **Enfoque Modular** Aunque ofrece adaptabilidad, la estructura de SEMMA puede no ser tan ampliamente aplicable en diversas situaciones.	**Moderada** **Diversas Etapas** Su flexibilidad se limita a las etapas del proceso y puede no ser tan granular como el enfoque de estructura modular de CRISP-DM.
Enfoque en el Negocio	Fuerte	Moderado	Moderado
Adopción en la Industria	Amplia	Moderada	Moderada

Tabla 1: Comparativa entre metodologías KDD, CRISP-DM y SEMMA [11]

En el caso específico de este Proyecto de Título se ha decidido utilizar una metodología CRISP-DM modificada, que excluirá la etapa de implementación. La elección de esta metodología se debe a la estructura clara, detallada y flexible de CRISP-DM, junto con su enfoque fuerte en la comprensión del negocio y su amplio uso en proyectos similares de la empresa, que la posiciona como una metodología que integra todos los aspectos necesarios para el desarrollo del proyecto.

7.3 MODELOS PREDITIVOS

Existen modelos que en la literatura han presentado un buen desempeño en aplicaciones similares, los cuales se describen a continuación.

Árboles de decisión

Los Árboles de decisión son estructuras con forma de árbol que representan decisiones capaces de generar reglas de clasificación para un conjunto de datos específico. La desventaja de este modelo, es que no son los más adecuados para capturar relaciones complejas y no lineales entre los atributos, las que generalmente se presentan al estudiar el comportamiento de fuga [39].

Redes Neuronales

Los modelos de redes neuronales son un enfoque popular para abordar problemas complejos, como lo es el fenómeno de la fuga de clientes. Demostrando un mejor rendimiento en comparación con los árboles de decisión [43].

Las redes neuronales ofrecen una serie de ventajas, que incluyen la capacidad de detectar de manera implícita relaciones complejas y no lineales entre variables dependientes e independientes, la capacidad de detectar todas las posibles interacciones entre las variables predictoras y la disponibilidad de múltiples algoritmos de entrenamiento. Sin embargo, tienen desventajas, como su naturaleza de "caja negra", que hace referencia a la baja interpretabilidad del modelo. Además, requiere de una gran carga computacional [43].

Análisis de Supervivencia

El análisis de supervivencia, o también referido como análisis de tiempo hasta un evento, se refiere a un conjunto de métodos estadísticos para analizar la duración del tiempo hasta la ocurrencia de un punto final bien definido de interés. Por lo general, no todos los individuos experimentan el evento (por ejemplo, la fuga) al final del período de observación, por lo que los tiempos de supervivencia reales de algunos individuos son desconocidos. Este fenómeno, conocido como censura, debe ser tenido en cuenta en el análisis para permitir inferencias válidas[44].

A continuación se presenta un tabla que compara los modelos mencionados en tres aspectos clave para el desarrollo del objetivo general del proyecto.

Aspecto	Análisis de Supervivencia	Bosque Aleatorio	Redes Neuronales Artificiales
Análisis de Tiempo hasta Evento	Modelo explícito del tiempo hasta que ocurre un evento.	No se enfoca inherentemente en el análisis de tiempo hasta el evento.	No se enfoca inherentemente en el análisis de tiempo hasta el evento.
Manejo de Datos Censurados	Diseñado específicamente para manejar datos censurados.	Puede manejar datos censurados en cierta medida.	Puede manejar datos censurados, pero puede requerir cuidados adicionales.
Interpretabilidad	Proporciona curvas de supervivencia interpretables y tasas de riesgo.	Ofrece importancia de características, indicando la contribución de las variables.	Menos interpretable debido a estructuras complejas y no lineales.

Tabla 2: Resumen tabla comparativa Árboles de decisión, Análisis de Supervivencia y Redes Neuronales. [1] [8] (Ver tabla completa, Tabla 2 en Anexos)

A partir del análisis comparativo de los modelos, se determinó que la técnica que mejor se adapta a los requerimientos del objetivo general corresponde al análisis de supervivencia, principalmente por su capacidad de incorporar datos censurados, los cuales corresponden a más de la mitad de los datos totales para este análisis.

Otra ventaja relevante del análisis de supervivencia respecto de los otros dos modelos, corresponde al aspecto temporal de la predicción del comportamiento de fuga. Esto es especialmente importante en el caso de la retención, donde se busca llevar a cabo acciones preventivas, con un enfoque proactivo sobre uno reactivo. La

temporalidad proactiva implica lanzar una campaña dirigida a clientes que han sido identificados como riesgo de pérdida, pero que aún no han abandonado. Por otro lado, la temporalidad reactiva se refiere a cuando la empresa intenta evitar que un cliente se dé de baja en el preciso momento en que ese cliente está a punto de hacerlo.

Por otro lado, este análisis permite una mejor escalabilidad para ampliar el análisis de permanencia, ya que permite incluir una mayor cantidad de datos que consideren el aspecto temporal, lo cual es primordial para un posterior análisis más detallado de fuga de los clientes que incluya variables comportamentales. Esta conciencia temporal permite una comprensión de la evolución de las interacciones con el cliente, facilitando la identificación de períodos críticos y la evaluación de riesgos cambiantes a lo largo de todo el ciclo de vida del cliente.

7.4 ANÁLISIS DE SUPERVIVENCIA

A continuación se detalla el modelo de Análisis de Supervivencia, cuya información se obtiene principalmente a partir de la documentación de la librería de Python, *lifelines* [47] y el libro Applied Survival Analysis: Regression Modelling of time to event data, escrito por David W. Hosmer, Stanley Lemeshow, Susanne May [46].

7.4.1 APLICACIONES

El Análisis de Supervivencia se desarrolló para evaluar las expectativas de vida de los individuos, siendo principalmente utilizado por actuarios y profesionales de la salud. El estudio podía abarcar desde la población de una nación en el caso de los actuarios hasta un grupo afectado por una enfermedad específica en el caso de los profesionales de la salud. [46]

No obstante, el análisis de supervivencia puede ampliarse más allá del ámbito de los eventos de nacimiento y muerte para estudiar el tiempo entre dos eventos definidos. En la literatura hay múltiples aplicaciones de análisis de supervivencia, por ejemplo: en el análisis del abandono de la educación superior [38], el estudio de la evolución temporal de la mejora bajo un tratamiento antidepresivo [40], el modelamiento del tiempo hasta renuncia para estimar los tiempos de recambio de personal en una empresa.

En el contexto del Trabajo de Título, el estudio se realiza respecto al tiempo que transcurre desde la apertura de una línea móvil postpago (nacimiento) hasta su desactivación o fuga de la línea móvil (muerte). Por simplicidad a partir de ahora, se hará referencia a estos eventos como ***habilitación*** y ***fuga*** respectivamente.

7.4.2 FUNCIÓN DE SUPERVIVENCIA

En el análisis de supervivencia, el tiempo de vida se modela como una variable aleatoria, que se denota como T, y que solo puede tomar valores positivos. Se dirá que T es un tiempo de vida aleatorio.

Sea T un tiempo de vida aleatorio no negativo tomado de la población de estudio. La función de supervivencia, $S(t)$ se define como la probabilidad de que el evento de interés (fuga) no haya ocurrido en el momento *t,* es decir:

$$S(t) = Pr\,(T > t)$$

Esta función cumple que 0 ≤ $S(t)$ ≤ 1 para todo *t* positivo. Por lo demás este modelo considera que el tiempo es continuo, por lo que se supondrá que las variables aleatorias T tendrán una función de distribución $f_T(t)$. Por último, la esperanza de la variable aleatoria *T*, que se denomina *E[T]*, se llamará esperanza de permanencia.

7.4.3 FUNCIÓN DE RIESGO

Un parámetro importante dentro del análisis de supervivencia es la función de riesgo que se define según la siguiente expresión:

$$h(t) = \lim_{\delta \to 0} \frac{Pr(t \leq T \leq t + \delta t \mid T > t)}{\delta t}$$

Esta expresión de alguna forma nos indica la “propensión” que tiene un cliente de fugarse al “instante siguiente”. Con esta magnitud se puede recuperar la función de supervivencia haciendo la siguiente observación:

$$h(t) = \frac{-S'(t)}{S(t)}$$

Y así claramente S(t) se expresa como la exponencial de algo que depende de $h(t)$.

$$S(t) = exp(-\int_0^t h(u)du)$$

En particular, si $h(t)$ es una función constante igual a λ entonces la variable aleatoria T seguiría la distribución más básica de los modelos de supervivencia, es decir, una exponencial de parámetro λ. De esta manera, $h(t)$ se interpreta como una forma de traducir la riqueza de comportamiento.

A su vez, la función de riesgo acumulada $H(t)$, cumple un rol similar a $h(t)$ y corresponde simplemente a la integral de 0 a t de $h(t)$. En particular, la función de supervivencia queda expresada de la siguiente forma:

$$S(t) = exp(-H(t))$$

7.4.4 ESTIMADORES PARAMÉTRICOS

Los modelos de supervivencia paramétricos son aquellos en los que la distribución de T posee una fórmula explícita y predeterminada para la cual simplemente hay que ajustar ciertas constantes que aparecen en dicha fórmula. La ventaja de estos modelos reside en la capacidad predictiva, ya que al asumir una distribución, los datos son fácilmente extrapolables. Entre los modelos paramétricos más comúnmente usados en este tipo de análisis, se encuentran el modelo exponencial y el modelo Weibull [35]

7.4.5 ESTIMADOR NO PARAMÉTRICO

Existe una gran variedad de estimadores para la función de supervivencia. En particular, el estimador Kaplan-meier se caracteriza por ser un modelo no paramétrico ampliamente utilizado [26]. A diferencia de los estimadores paramétricos, un estimador no paramétrico se ajusta a la curva sin preestablecer una distribución de probabilidad, y por lo mismo tiene un enfoque descriptivo al sólo considerar los datos existentes.

Para definir el estimador, es necesario recordar que el tiempo estará discretizado de forma homogénea y que cada instante de tiempo se denota como t_i. Con ello, n_i es el número de observaciones en riesgo de muerte y d_i el número de muertes observadas a tiempo t_i. Con ello, el estimador Kaplan-Meier en tiempo t, se define como el siguiente producto:

$$\widehat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

7.4.6 MODELO COX PROPORTIONAL HAZARD

El modelo Cox Proportional Hazard corresponde a una técnica estadística que permite estudiar el efecto simultáneo de una o más variables en la distribución de supervivencia. Se define de la siguiente manera

$$h(t|X) = h_0(t)\, exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

- $h(t)$ corresponde al riesgo esperado en el tiempo *t*
- $h_0(t)$ Es el riesgo base, es decir cuando todas las covariables se igualan a cero.
- $\beta_1, \ldots, \beta_p$ Son los parámetros del modelo que describen el efecto de las variables.

Las variables con coeficientes positivos (los valores de β) están asociadas con un aumento en el riesgo y una disminución en la duración de la supervivencia, es decir, a medida que el predictor aumenta, el riesgo del evento aumenta y la duración de supervivencia prevista disminuye. Los coeficientes negativos indican un menor riesgo y una mayor duración de supervivencia.

8. METODOLOGÍA

A continuación se describen las etapas de la metodología utilizada para la realización del proyecto, cuya selección fue explicada en el Capítulo 7, Marco Teórico. La información presentada a continuación se obtuvo a partir del sitio web de la empresa tecnológica internacional, IBM.

8.1 COMPRENSIÓN DEL NEGOCIO

Comprender el funcionamiento del negocio y la industria de telecomunicaciones, con especial foco en el mercado móvil en Chile. Además, se debe comprender en profundidad las necesidades del área para la cual se desarrolla el proyecto y definir los objetivos y alcances del proyecto en base a esto.

8.2 COMPRENSIÓN DE LOS DATOS

En esta sección se lleva a cabo una exploración de los datos disponibles en la empresa que sean útiles para el desarrollo del proyecto. En esta etapa se debe determinar qué datos no se tienen y qué datos son posibles de conseguir, explorando las tablas de información disponibles y manteniendo comunicación con otros equipos dentro de la empresa que podrían tener conocimiento o acceso a otras tablas de utilidad para el modelo. Entender los datos con los que se dispone y la calidad de estos y comunicar los resultados que se esperan lograr y los posibles problemas que se pueden generar respecto a la idea inicial de proyecto propuesta.

8.3 PREPARACIÓN DE LOS DATOS

Esta etapa es una de las más relevantes dentro de la metodología y, en general, en la que se invierte la mayor cantidad de tiempo. Consiste en la fusión de conjuntos y registros de datos, selección de una muestra o subconjunto de datos, agregación de registros y creación de nuevas variables o atributos.

8.3.1 SELECCIÓN DE DATOS

Luego de preparar los datos, se deben seleccionar aquellos que son relevantes para los objetivos del proyecto. Para ello, se realiza la selección de variables. Las variables representan atributos o características de cada observación. Por otro lado también se pueden filtrar las observaciones, por ejemplo para filtrar por criterios específicos, establecer condiciones temporales o geográficas, entre otras.

8.3.2 LIMPIEZA DE DATOS

Posteriormente se debe realizar una inspección más profunda de la calidad de los datos previamente seleccionados, además de buscar soluciones para aquellos registros que presenten problemas. Los problemas más comunes son los valores faltantes y los valores incoherentes. En cada caso se debe evaluar cómo proseguir, ya sea excluyendo aquellas filas del análisis o realizando una estimación del valor. Esto dependerá de la cantidad de datos que presenten este problema, y de la información disponible para realizar estimaciones.

8.3.3 CONSTRUCCIÓN DE NUEVOS DATOS

Utilizando la información existente es posible generar nuevas variables que sean relevantes para el fenómeno de estudio y que permitan mejorar el rendimiento del modelo. En el caso específico de este proyecto, la creación de nuevas variables serán útiles para mejorar la capacidad predictiva del modelo. [20]

8.3.4 INTEGRACIÓN DATOS

Los datos necesarios para el desarrollo del proyecto provienen de distintas tablas, cuya información relevante debe ser consolidada en una base final que contenga toda la información necesaria para el modelo.

Esta integración se realiza mediante la fusión o adición de datos. En primer lugar, la fusión de datos implica unir dos conjuntos de datos con registros similares, pero con atributos diferentes. Los datos se fusionan utilizando el mismo identificador clave en cada registro y las tablas resultantes aumentan las columnas o las características. Por otro lado, la adición de datos implica integrar dos o más conjuntos de datos con atributos similares, pero con registros diferentes.

8.3.5 FORMATO DE DATOS

Previo al modelamiento, es fundamental adaptar los datos al formato que admite el modelo. Existen técnicas que realizan automáticamente las transformaciones de los datos por lo que no es necesario hacerlo previamente. Sin embargo, hay algoritmos

que admiten formatos específicos para los datos de entrada y, en estos casos, se debe adaptar la base consolidada al formato que corresponda.

8.4 MODELADO

Una vez consolidados los datos que se utilizarán en el modelo, se deben realizar múltiples iteraciones del algoritmo de la técnica seleccionada, probando en cada repetición distintas combinaciones de variables y parámetros con el objetivo de encontrar aquella que optimice el rendimiento del modelo y entregue los resultados más precisos. Para esto se deben definir métricas de evaluación que permitan evaluar y comparar los modelos en cada iteración.

En el caso específico de este proyecto, el modelo se evaluará principalmente a través del indicador *concordance,* que corresponde a un índice que evalúa la capacidad predictiva del modelo de regresión de Cox en el análisis de supervivencia. Su valor varía entre el 0 y 1 y su interpretación es la siguiente [27]:

- Un índice de concordancia de 0.5 indica que el modelo no tiene capacidad predictiva y se desempeña tan bien como un lanzamiento de moneda al predecir eventos.

- Un índice de concordancia de 1.0 indica que el modelo predice perfectamente quién experimentará el evento antes que otros.

En síntesis, un modelo con mayor *concordancia* tiene una mayor capacidad predictiva.

8.5 TRASPASO DEL MODELO

Finalmente, dentro de la metodología se incluye el traspaso del modelo. El primer paso de esta etapa final consiste en el respaldo de todos los códigos y archivos asociados al proyecto en Bitbucket, que corresponde a una herramienta de alojamiento de código y colaboración diseñada para equipos [4], ampliamente utilizada en Entel para almacenar los proyectos de Data Science. Además, se realizará una capacitación en la que un integrante del equipo deberá ejecutar el modelo y obtener las predicciones para el mes de noviembre.

9. DESARROLLO

9.1 COMPRENSIÓN DEL NEGOCIO

Esta primera etapa se llevó a cabo junto a Paz Obrecht, tutora de la alumna en la empresa y jefa del equipo de Analytics CVM. En esta instancia se conversó sobre el uso que se le dará a la información extraída a partir del modelo, para entender en profundidad las necesidades del área.

En esta etapa se abordó el modelo actual relacionado a la fuga de clientes y sus limitaciones, las cuales se relacionan con el aspecto temporal de la predicción, ya que el modelo predice únicamente la tasa de fuga del mes siguiente. Por otro lado, la información de dicho modelo no es lo suficientemente precisa, es decir se entrega un porcentaje de fuga respecto a la base total, el modelo no especifica los clientes con mayor riesgo.

A partir de estos dos factores fundamentales para el equipo de CVM, se determinó que el modelo debe entregar información respecto al tiempo de permanencia de las líneas móviles a través del tiempo. Por otro lado, la información debe permitir identificar el riesgo de fuga a nivel de línea móvil, para que las campañas apunten únicamente a aquellos clientes con intención de fuga.

Por otro lado, se determinó la forma en la que se traspasará la información al equipo. Según lo acordado, el modelo deberá estimar la permanencia estimada para cada cliente, además de su probabilidad de fuga en tiempos críticos, que se establecieron como: 3,6,9 y 12 meses luego de la habilitación.

9.2 COMPRENSIÓN DE LOS DATOS

Las fuentes de información que se consideraron para el modelo, están almacenadas en Oracle, herramienta para la gestión de Bases de Datos que se usa principalmente en grandes empresas, diseñado para que las organizaciones puedan controlar y gestionar grandes volúmenes de datos [16]. Entel almacena gran parte de sus datos en esta plataforma, de la cual se extrajo información de 3 tablas, detalladas a continuación.

En primer lugar, la base ‘habilitaciones’ contiene toda la información respecto a la línea móvil y al cliente titular al momento de la habilitación. Por otro lado, se extrajo información de una tabla ‘fuga’ que contiene información relacionada a la línea móvil y cliente titular al momento de la fuga. Finalmente, se consideró la información de la tabla ‘Malla Parental’ que contiene información de los clientes titulares de las líneas.

Debido a que el modelo se utilizará para la permanencia de nuevos clientes, sólo puede incluir variables que describen sus características al momento de habilitar la línea, como lo son: edad, estado civil, origen del cliente, forma de pago, número de líneas asociadas al titular, etc. Por esta razón, la mayor parte de la información proviene de la tabla de habilitaciones, y no se ha considerado necesario incluir información de otras bases, por ejemplo: calidad de red, satisfacción del cliente, tráfico de datos, etc.

9.3 PREPARACIÓN DE LOS DATOS

9.3.1 SELECCIÓN DE LOS DATOS

En primer lugar, se seleccionaron aquellos datos que luego permiten determinar, si el evento de fuga ha ocurrido o no, y luego la cantidad de meses que ha transcurrido desde la habilitación hasta la fuga para aquellos que cancelaron el servicio, o hasta la actualidad en el caso de las líneas que siguen vigentes. Ambas variables son indispensables en el análisis de supervivencia y se obtienen a partir de las fechas de habilitación y fechas de fuga.

Para la selección del resto de las variables, la única condición que se tuvo en cuenta fue que la información estuviera disponible al momento de la habilitación. Como todas las variables de la tabla habilitaciones cumplen con esa condición se consideró la base entera. Lo mismo en el caso de la tabla Malla Parental. En cambio, en la tabla fuga sólo se consideró la variable que indica el periodo de fuga. A continuación se presenta una tabla que describe las variables más relevantes. La tabla con todas las variables se encuentra disponible en la Tabla 4 de Anexos.

Nombre Variable	Descripción	Formato	Valores
Periodo	Fecha de habilitación de la línea	número entero (yyyymm)	Se consideran las fechas desde julio de 2021 hasta la actualidad
Tipo Actividad Comercial	Describe el tipo de habilitación	cadena de caracteres (VARCHAR)	Migración prepago a postpago Port in: línea portada desde otra compañía Habilitación: Línea nueva (no existía previamente)
Tipo de línea	Indica si la línea está asociada a un cliente de la empresa o es un nuevo cliente	VARCHAR	Additional: corresponde a una línea adicional Base: primera línea habilitada
Precio de suscripción	Representa el precio del plan contratado	Valor entero	Valores entre 0 y 69.990
Canal de venta	Especifica sobre el tipo de canal	Cadena de caracteres	S2S, ONLINE PÚBLICO, TIENDAS, OUTBOUND, IVR,GGTT, MASIVOS,

Tabla 3: Describe algunas de las variables seleccionadas, elaboración propia.

9.3.2 LIMPIEZA DE DATOS

Una vez seleccionados los datos, se llevó a cabo la primera exploración para determinar el tipo de variable y la cantidad de datos nulos o desconocidos en cada una de las tablas. Una vez seleccionados los datos, fueron exportados a CSV para luego ser procesados en Python.

Durante esta etapa se ajustaron los formatos de las variables correspondientes y se realizó una exploración de los valores nulos y valores outliers en cada variable. A partir de la exploración se tomó la decisión de eliminar la variable descuento debido a que la gran mayoría de sus valores (cerca del 90%) correspondían a valores catalogados como 'DESCONOCIDO'. Los demás datos se mantuvieron al no presentar nulos ni outliers.

9.3.3 CONSTRUCCIÓN DE NUEVAS VARIABLES

La construcción de nuevas variables se realizó a partir de las variables previamente seleccionadas, y consistió principalmente en la agrupación de valores en intervalos con el fin de obtener el comportamiento de segmentos y así facilitar la interpretación y mejorar la eficiencia computacional.

Variable	Descripción	Variable de origen	Valores
Segmento edad	Las edades se dividen en 4 intervalos según el intervalo al que pertenezcan, los que se definieron según percentil 25, 50 y 75.	Edad_titular	1) 0-32 2) 32-41 3) 41-43 4) 53+
Trimestre	Trimestre en el que se realiza la habilitación	Periodo	T1: enero, febrero y marzo T2: abril, mayo y junio T3: julio, agosto y setiembre T4: octubre, noviembre y diciembre
gestión canal	Se agrupan los valores originales de la variable en los principales canales	ges_grupo_Canal	S2S, ONLINE, TIENDAS, OUTBOUND, DESCONOCIDO,OTRO

Tabla 4: Descripción de nuevas variables. Fuente: Elaboración propia

9.3.4 INTEGRACIÓN DE DATOS

Luego de las transformaciones anteriores, se eliminaron las variables originales, es decir, edad del titular, periodo y la modalidad de venta, para ser reemplazadas por los segmentos.

9.3.5 FORMATO DE DATOS

Los modelos Kaplan-Meier como el Cox-Proportional Hazard admiten datos del tipo numérico o binarios, por esta razón se debe realizar una codificación de las variables categóricas, proceso que consiste en convertir las variables categóricas a numéricas. En esta primera etapa del proyecto se realizó una codificación One-hot, que crea una columna binaria por cada valor único dentro de las variables.

9.3.5 EVALUACIÓN

Para la evaluación, se clasificaron las observaciones según el decil asociado a la predicción de la esperanza de permanencia. Luego, para cada uno de los 10 segmentos, se estimó la curva de supervivencia a través del estimador Kaplan Meier a partir de los datos reales de permanencia de cada una de las observaciones.

10. RESULTADOS

Los primeros resultados se obtuvieron a partir de la estimación de la curva de supervivencia para toda la base de clientes a través del estimador no paramétrico, Kaplan Meier, cuya gráfica se puede ver en el siguiente gráfico, en la que, eje Y representa la proporción de observaciones que aún no se fuga en tiempo t, determinado por el eje X.

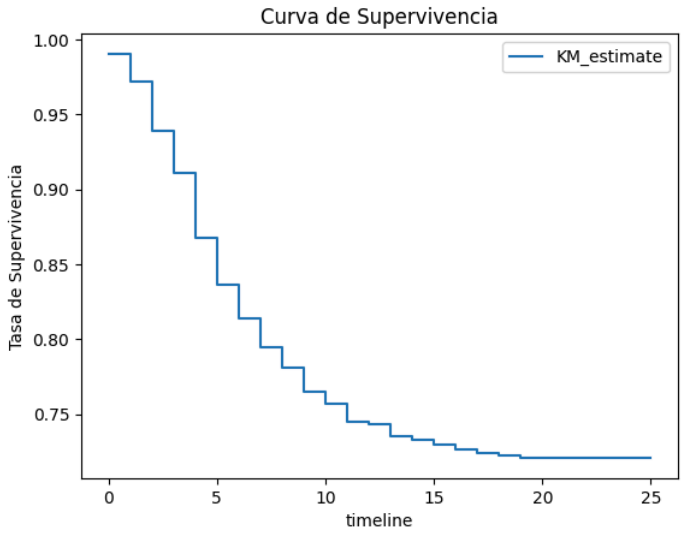

Figura 3: Curva de Supervivencia para base de clientes al momento de su habilitación.

Lo primero que se puede destacar del gráfico es que la tasa de supervivencia se mantiene por sobre 0,7 en todo el periodo de tiempo estudiado. Observando más detalladamente, se constata una alta tasa de fuga durante los primeros 5 meses, donde la curva presenta una pendiente con un mayor grado de inclinación. Posteriormente, la pendiente disminuye entre los 5 y 10 meses y de forma más notoria entre los 10 y 15 meses. Finalmente la curva se estabiliza a partir del mes número 15.

Lo anterior indica que el mayor riesgo de fuga ocurre los primeros meses, especialmente entre los 0 y 5 meses, donde la probabilidad de supervivencia cae con mayor intensidad, en 0,2. Es fundamental identificar aquellos clientes con mayor propensión a la fuga esos primeros meses para desplegar las estrategias de retención correspondientes.

El segundo análisis se realizó estimando la curva supervivencia para cada una de las variables categóricas (tipo de actividad comercial, tipo de línea, tipo de línea adicional, segmento de precio de destino, trimestre y segmento de número de líneas asociadas al rut titular). Esta vez se describe la curva de supervivencia para cada segmento dentro de la variable de estudio. Dichas gráficas se muestran a continuación.

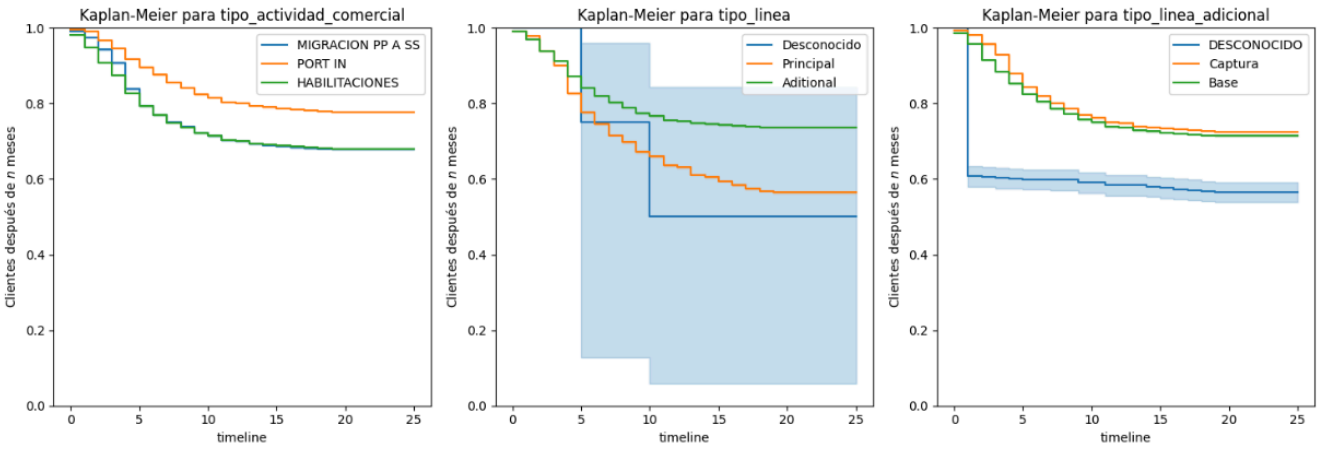

Figura 4,5 y 6: Curvas de supervivencia para variable Actividad Comercial, Tipo Línea y Tipo línea Adicional

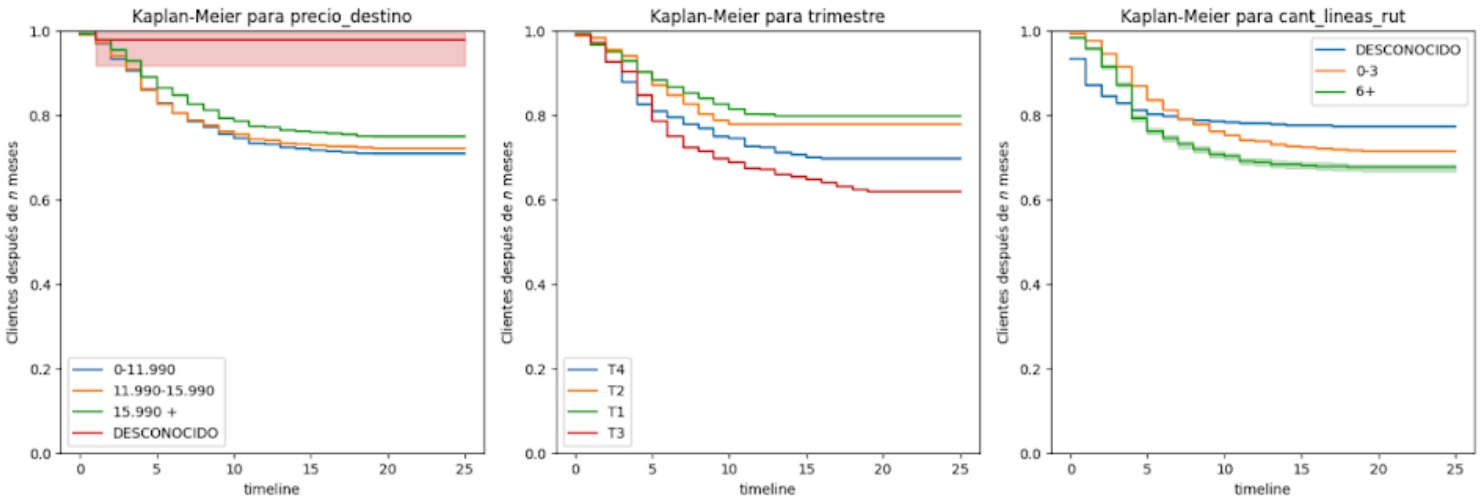

Figura 7,8 y 9: Curvas de supervivencia para variable Precio destino, Trimestre y Cantidad de líneas por RUT.

Para este análisis se consideraron las variables tipo de actividad comercial, fibra, forma de pago, gestión canal de venta, tipo de línea, tramo etario, número de integrantes en el hogar, estado civil, precio destino, trimestre, número de líneas por RUT. A continuación se describe la supervivencia de aquellas variables en las que existe una diferencia significativa en la probabilidad de supervivencia de las distintas categorías que la componen, es decir, cuando existe una diferencia >= 0.1.

- *tipo de línea:* en esta variable también se puede identificar con claridad una diferencia entre ambas categorías de la variable. La categoría adicional tiene una pendiente menor los primeros 4 meses, mientras que la categoría principal tiene una pendiente que disminuye recién al mes 20. Las probabilidades de supervivencia al mes 25 son de 0.75 y 0.68 para las líneas principales y adicionales, respectivamente.

- *Forma de pago:* Esta es una de las variables que mostró un resultado más significativo en la supervivencia, específicamente aquellos clientes que están sujetos a un pago automático (categoría 11 en la gráfica de anexos) tienden a tener una probabilidad de supervivencia mayor en 0.2 con respecto a aquellos clientes cuyo pago no se realiza de forma automática.
- *Tramo edad:* Al juntar tramos etarios en los siguientes grupos: 0-32, 32-41 y 53+ años, se logró observar una similitud entre los últimos dos tramos que, a su vez, se diferencian del primer tramo con una probabilidad de supervivencia mayor en 0.1 a lo largo del tiempo.

- *trimestre:* esta variable presenta una probabilidad superior en 0.2 en la supervivencia de aquellas líneas que se habilitaron el primer trimestre con respecto a las líneas habilitadas en el tercer trimestre.

Después del análisis por variable, se estudió el efecto simultáneo de todas las variables de la base consolidada en la supervivencia, para lo cual se implementó un modelo Cox-Proportional Hazard, que obtuvo un valor de 0.66 en el indicador concordance.

El valor obtenido en el indicador concordance es bueno, ya que indica que el modelo está logrando predecir la permanencia con un 0.66 de probabilidad. Más aún, en Entel se considera que un modelo tiene un buen rendimiento, si tiene una capacidad de predicción mayor a 0.6.

En la Tabla 3 de Anexos se muestra la tabla completa con el resumen del modelo. A continuación se presentan las variables con los coeficientes de mayor magnitud y que, por ende, tienen mayor influencia en la probabilidad de supervivencia.

Variable	Coeficiente
forma_pago_13	0.29
trimestre_T2	-0.06
trimestre_T3	0.39
trimestre_T4	0.19
lineas_rut_6+	0.17
actividad_PORTIN	-0.33
precio_15.990+	0.1
gestión_S2S	0.28

Tabla 5: Resumen coeficientes de variables que más impactan al riesgo de fuga, obtenidos a través del modelo Cox Proportional Hazard.

A partir de la tabla anterior, se puede concluir que la variable que más aumenta el riesgo de fuga es la variable trimestre_T3, ya que presenta un coeficiente positivo de 0.39, lo cual se interpreta como que dicha variable tiene probabilidades de fuga para aquellos clientes que habilitan la línea en el último trimestre. También, una variable que se asocia a un alto riesgo de fuga, es la variable de forma_pago_13, cuyo coeficiente tiene un valor de 0.29.

Por otro lado, existen variables con coeficiente negativo, y que presentan un menor riesgo a la fuga. La variable que presenta el menor riesgo es la actividad port in, que corresponde a todas aquellas líneas que se portan desde otra compañía. Otra variable que se asocia a un menor riesgo de fuga, aunque con menor impacto, es la variable que representa las habilitaciones en el segundo trimestre.

Sumado a lo anterior, el modelo también entrega información a nivel de cada observación, es decir, a nivel de línea móvil. Más precisamente, se estima la esperanza de permanencia de cada línea, además de la probabilidad de fuga en meses específicos: 3, 6, 9 y 12 meses a partir de la habilitación.

Una forma usual de evaluar los modelos predictivos en Entel es observando si este permite ordenar por decil. Más precisamente, se estimó la esperanza de permanencia para cada una de las líneas móviles. A cada valor predicho se le asignó el decil correspondiente. Este valor es incorporado como una etiqueta que, a priori, debería indicar un menor riesgo de fuga a medida que crece su valor. Para verificar lo anterior se estimaron las curvas de supervivencia en cada decil a través del estimador Kaplan-Meier. Cabe destacar que para obtener estas curvas no se utilizaron las predicciones, sino que los valores reales y los deciles recién designados. A continuación, el siguiente gráfico muestra el comportamiento de cada uno de los deciles.

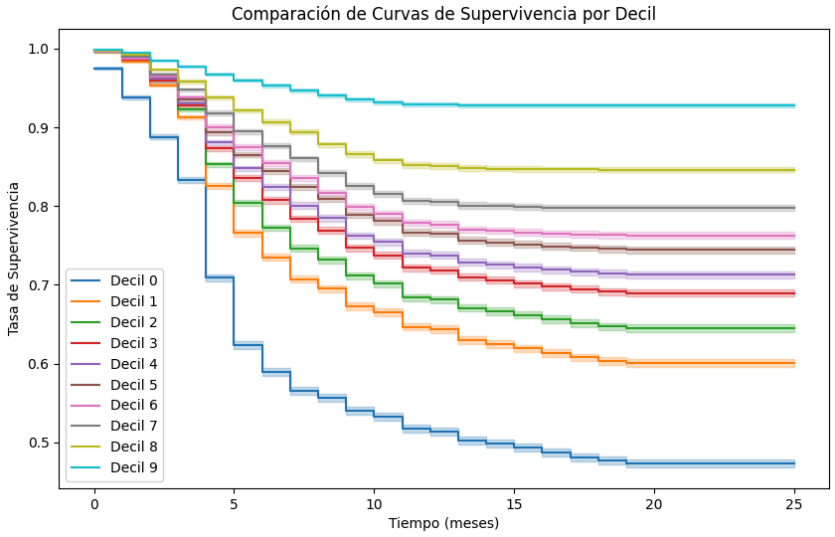

Figura 10: Evaluación del modelo predictivo.

A partir del gráfico se comprueba que las observaciones logran ordenar por decil. Esto pues las curvas no presentan cruce, y mantienen el orden a lo largo del tiempo, donde la curva del decil *i* se mantiene por debajo de la curva del decil *i+1.* Sumado a lo anterior, el índice de concordancia obtenido es de 0.66 lo cual cumple los estándares de los modelos del equipo de Analytics, cuyos modelos alcanzan, en su mayoría, valores entre 0.6 y 0.7 para este indicador.

11. CONCLUSIONES

Debido a la naturaleza iterativa de la metodología CRISP-DM, se pudo mejorar progresivamente el Modelo de Supervivencia al realizar, por ejemplo, ajustes en los datos. Las iteraciones se enfocaron, principalmente, en el procesamiento y limpieza de datos y en la inclusión de nuevas variables al modelo. Como resultado, el modelo final permite predecir la permanencia de nuevos clientes y determinar aquellos clientes que presentan mayor riesgo de fuga en tiempos determinados. Finalmente, se llevó a cabo el traspaso del modelo. Así, es posible afirmar que los objetivos específicos se cumplieron, y en consecuencia, el objetivo general.

Durante todo el desarrollo del proyecto fue fundamental el apoyo de la tutora de la alumna en la empresa, Paz Obrecht. Paz, desde su experiencia, se encargó de supervisar el proyecto y entregar feedback que fue incorporado regularmente. Además, en una primera instancia, el apoyo por parte del equipo de Analytics fue clave para el manejo adecuado de las herramientas y plataformas propias de Entel.

Otro aspecto fundamental que permitió el cumplimiento del objetivo fue la disponibilidad de una gran cantidad de datos almacenados en la empresa, los que fueron incorporados como variables al modelo. Los datos históricos son necesarios para identificar patrones y relaciones entre las variables, y mientras más información exista, mejor el modelo puede captarlos, lo que se traduce en una mejor capacidad predictiva.

La preparación de los datos fue uno de los grandes desafíos del proyecto, ya que fue necesario depurarlos en múltiples ocasiones. En efecto, se invirtió más tiempo de lo esperado en aquello. Esta etapa consistió principalmente en: la creación de variables, detección y tratamiento de datos faltantes e identificación y corrección de *outliers*. Para ello fue fundamental el conocimiento y manejo de las herramientas y plataformas utilizadas por la empresa, como Python y SQL.

Los resultados obtenidos cumplen los requisitos definidos al comienzo del proyecto y la información que genera el modelo podría ser útil para el que el equipo de CVM desarrolle estrategias de retención o fidelización. Gracias al modelo se obtiene: la esperanza de la permanencia de cada línea móvil, el decil asociado a la continuidad de la línea y las variables destacadas que más afectan (ya sea positiva o negativamente) en la permanencia.

En primer lugar, la esperanza de la permanencia asociada a cada línea permite llevar a cabo valorizaciones, es decir, aproximar el monto en dinero que aportará cada una de ellas (multiplicando dicha esperanza con el precio pagado mensualmente por la línea). Por otro lado, este indicador también permite identificar aquellos clientes de alta permanencia respecto a aquellos de baja permanencia. Sin embargo, para dicha identificación se recomienda utilizar la categoría de deciles, ya que las líneas al interior de cada decil presentan un comportamiento similar.

Respecto a las variables destacadas, se recomienda potenciar las ventas a través de aquellos factores asociados a líneas con mayor permanencia. Por ejemplo, como se mostró en la sección de resultados, los clientes habilitados por Port In tienden a mantenerse más tiempo en la empresa. De esta forma, se propone aumentar los esfuerzos comerciales y de marketing en captar clientes de otras compañías. Una conclusión similar se puede realizar con respecto a las otras variables expuestas en dicha sección.

Por último, es importante mencionar que, si bien este proyecto es un primer acercamiento al comportamiento de fuga en la empresa, se pueden desarrollar modelos más complejos cuya predicción permita entender la fuga de clientes. En particular, para un siguiente paso se propone generar un modelo que estudie toda la trayectoria del cliente. Este análisis permitiría incluir variables como: la cantidad de tráfico de datos, número de reclamos, calidad de la señal, entre otras. Este enfoque permitirá enriquecer la información sobre los clientes y llevar a cabo estrategias de retención y fidelización que apunte al stock de clientes, el segmento mayoritario.

12. BIBLIOGRAFÍA

[1] Alberts, L., Bsc, Peeters, R.L., Braekers, R., Meijer, C., & Netherlands, V. 2006. Churn Prediction in the Mobile Telecommunications Industry An application of Survival Analysis in Data Mining Master Thesis.

[2] Banco Central 2022. CUENTAS NACIONALES DE CHILE Métodos y Fuentes de Información.

[3] Biblioteca del Congreso Nacional de Chile, Abril 2020. Portabilidad numérica en Chile.

[4] BitBucket. Breve presentación de BitBucket https://bitbucket.org/product/es/guides/getting-started/overview#a-brief-overview-of-bitbucket

[5] Bnamericas, Mayo 2023. Entel: alzas en ingresos por servicios móviles y clientes con servicios sobre fibra óptica destacan en resultados del primer trimestre.

[6] Banco Mundial, Octubre 2023. Chile Panorama general. https://www.bancomundial.org/es/country/chile/overview [Consulta: 11 Novimebre 2023]

[7] Diario Financiero, Enero 2023. Chile cierra 2022 con la mayor inflación desde 1991 tras tras un IPC en diciembre que se ubicó dentro de lo esperado por el mercado. https://www.df.cl/economia-y-politica/macro/chile-cierra-2022-con-la-mayor-inflacion-desde-1991-tras-un-ipc-en [Consulta: 11 Novimebre 2023]

[8] Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. Journal of Multidisciplinary Developments.

[8] Challenger, Díaz, Becerra, Junio 2014. El lenguaje de programación Python.

[9] Chandar and Krishna, 2006. Modeling churn behavior of bank customers using predictive data mining techniques.

[10] Cuentas Nacionales de Chile, Métodos y fuentes de información. Banco Central de Chile 2022.

[11] Dåderman, A., & Rosander, S. (2018). Evaluating Frameworks for Implementing Machine Learning in Signal Processing : A Comparative Study of CRISP-DM, SEMMA and KDD.

[11] Diario Financiero, Septiembre 2023. Conexiones 5G se disparan en Chile y se acercan a los 3 millones de usuarios a nivel nacional. https://www.df.cl/empresas/telecom-tecnologia/conexiones-5g-se-acercan-casi-a-3-millones-de-usuarios-a-nivel-nacional

[12] Diario Estrategia, Agosto 2023. Entel: alza en ingresos por servicios fijos y móviles compensa efecto de menor actividad económica en el primer semestre

[13] Diario Financiero, Octubre 2023. Chile en la vanguardia de la conectividad que hay detrás del explosivo crecimiento de la red 5G. https://diariofinanciero.pressreader.com/article/281749863987217

[14] Entel, Información Corporativa Entel. https://informacioncorporativa.entel.cl/aniversario-entel

[15] Entel, Mayo 2023. Política Ambiental Entel

[16] Entel, 2022. Información Corporativa. https://informacioncorporativa.entel.cl/nuestra-compa%C3%B1ia

[17] Entel, 2022. Memoria Integrada 2022. https://entel.modyocdn.com/uploads/3be5d1be-f63c-4eea-a898-280596be8435/original/2304 11_Entel_Memoria_2023_Libro_web.pdf

[18] Gobierno de Chile, Mayo 2022. Lanzamos el Plan Brecha Digital Cero para que todas y todos tengan acceso a conectividad independiente del lugar en que viven. https://www.gob.cl/noticias/lanzamos-el-plan-brecha-digital-cero-para-que-todas-y-todos-tengan-acceso-conectividad-independiente-del-lugar-en-que-viven/

[19] Gury, 2011. Dropping out of higher education in France: a micro-economic approach using survival analysis, Education Economics.

[20] IBM, Agosto 2021. CRISP-DM de IBM SPSS Modeler. https://www.ibm.com/docs/en/spss-modeler/saas?topic=overview-crisp-dm-in-spss-modeler

[21] IBM, Agosto 2021. Limpieza de datos. https://www.ibm.com/docs/es/spss-modeler/saas?topic=preparation-cleaning-data

[22] IBM, Agosto 2021. Integración de datos. https://www.ibm.com/docs/es/spss-modeler/saas?topic=preparation-integrating-data

[23] IBM, Agosto 2021. Selección de datos.https://www.ibm.com/docs/es/spss-modeler/18.1.1?topic=preparation-selecting-data

[24] IBM, Agosto 2021. Formato de datos. https://www.ibm.com/docs/es/spss-modeler/saas?topic=preparation-formatting-data

[25] IBM, Agosto 2021. Generación de los modelos https://www.ibm.com/docs/es/spss-modeler/saas?topic=modeling-building-models

[26] Jager, van Dijk, Zoccali, Dekker, 2008. The analysis of survival data: the Kaplan–Meier method.

[27] Lambin, Raykar, 2007. On Ranking in Survival Analysis: Bounds on the Concordance Index

[28] Ministerio de Transporte y Telecomunicaciones. Ley general de telecomunicaciones.https://www.bcn.cl/leychile/navegar?i=29591&f=2020-11-26

[29] Ministerio de Transportes y Telecomunicaciones, Subsecretaría de Telecomunicaciones, Julio 2023. Reporte de portabilidad mensual.

[30] Ministerio de Transportes y Telecomunicaciones, Subsecretaría de Telecomunicaciones, Mayo 2023. Sector Telecomunicaciones Primer Trimestre 2023.

[31] Ministerio de Transportes y Telecomunicaciones, Subsecretaría de Telecomunicaciones, 2022. Sector de Telecomunicaciones cierre 2022. Año 2022.https://www.subtel.gob.cl/wp-content/uploads/2023/03/PPT_Series_DICIEMBRE_2022_V0.pdf

[32] Ministerio de Transportes y Telecomunicaciones, Subsecretaría de Telecomunicaciones, Mayo 2022. Especial Análisis Nueva Tecnología 5G en Internet

Móvil y crecimiento Tecnología Fibra en Internet Fija.

[33] Reichheld y Teal, Harvard Business School Press, 1996. The Loyalty Effect the Hidden Force behind Growth, Profits, and Lasting Value.

[34] Roberts, Rojas, Rojas, Agosto 2023. Estrategia de transformación digital Chile Digital 2035.

[35] Rodríguez, 2001. Parametric Survival Models.

[36] S&P Global Ratings, Agosto 2023. Intensa competencia e inversiones afectan los indicadores de operadores de telecomunicaciones en Chile en medio de esfuerzos por impulsar rendimientos.

[37] S & P Global Ratings, Junio 2023.Análisis Detallado Empresa Nacional de Telecomunicaciones S.A.

[38] Somers, 1996. Modelling employee withdrawal behaviour over time: A study of turnover using survival analysis.

[39] Song, Lu, 2015. Decision tree methods: applications for classification and prediction.

[40] Stassen, Delini-Stula, Angst, 1993. Time course of improvement under antidepressant treatment: A survival-analytical approach.

[41] Statista, Julio 2023. Distribución del producto interno bruto (PIB) por actividad económica en Chile en 2022.https://es.statista.com/estadisticas/1285944/participacion-de-las-actividades-economicas-en-el-pib-de-chile/

[42] Statista 2022. El despliegue de la 5G en el mundo. https://es.statista.com/grafico/23241/nivel-de-desarrollo-de-la-tecnologia-5g-en-el-mundo/#:~:text=La%205G%20se%20est%C3%A1%20extendiendo,38%20de%20mediados%20de%202020

[43] Tu, 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes

[44] Watt, Aitchison, MacKie, y Sirel, Octubre 1996. Survival analysis: the importance

of censored observations.

[45] ChatGPT

[46] David W. Hosmer, Stanley Lemeshow, Susanne May 2008. Applied Survival

[47] Cam Davidson-Pilon.2023. Lifelines Documentation

[48] Oxford, Personalización. https://www.oxfordlearnersdictionaries.com/definition/english/personalized

[49] Shobhana Chandra, Sanjeev Verma, Weng Marc Lim, Satish Kumar, Naveen Donthu.2022. Personalization in personalized marketing: Trends and ways forward

13. ANEXOS

Anexo A: Tabla Resumen modelo Cox Proportional Hazard

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
sexo_masculino_rut	0.02	1.02	0.00	0.02	0.03	1.02	1.03	0.00	10.43	<0.005	82.27
sexo_null_rut	0.04	1.04	0.00	0.03	0.05	1.03	1.05	0.00	9.57	<0.005	69.65
act__MIGRACION PP A SS	-0.02	0.98	0.00	-0.03	-0.02	0.97	0.98	0.00	-7.69	<0.005	46.00
act__PORT IN	-0.33	0.72	0.00	-0.33	-0.33	0.72	0.72	0.00	-136.50	<0.005	inf
canal_Desconocido	0.52	1.68	0.96	-1.36	2.39	0.26	10.91	0.00	0.54	0.59	0.77
canal_Principal	0.30	1.35	0.00	0.29	0.30	1.34	1.36	0.00	88.19	<0.005	inf
linea_adicional_Captura	-0.03	0.97	0.00	-0.03	-0.02	0.97	0.98	0.00	-10.62	<0.005	85.11
linea_adicional_DESCONOCIDO	0.29	1.33	0.04	0.20	0.37	1.22	1.45	0.00	6.52	<0.005	33.75
ges_grupo__ONLINE PRIVADO	-0.22	0.80	0.00	-0.23	-0.21	0.79	0.81	0.00	-51.26	<0.005	inf
ges_grupo__ONLINE PUBLICO	-0.17	0.85	0.00	-0.17	-0.16	0.84	0.85	0.00	-47.07	<0.005	inf
ges_grupo__OTRO	0.02	1.02	0.00	0.01	0.02	1.01	1.02	0.00	3.26	<0.005	9.80
ges_grupo__OUTBOUND	0.06	1.06	0.00	0.05	0.06	1.05	1.06	0.00	16.31	<0.005	196.29
ges_grupo__S2S	0.28	1.32	0.00	0.27	0.29	1.31	1.34	0.00	63.95	<0.005	inf
ges_grupo__TIENDAS	-0.05	0.95	0.00	-0.05	-0.04	0.95	0.96	0.00	-13.77	<0.005	140.90
titular_1.0	-0.02	0.98	0.00	-0.02	-0.01	0.98	0.99	0.00	-7.19	<0.005	40.51
titular_DESCONOCIDO	-0.82	0.44	0.01	-0.84	-0.81	0.43	0.45	0.00	-98.58	<0.005	inf
fibra__1.0	0.00	1.00	0.00	-0.01	0.01	0.99	1.01	0.00	0.53	0.60	0.74
fibra__DESCONOCIDO	0.68	1.98	0.01	0.67	0.70	1.95	2.01	0.00	94.58	<0.005	inf
fomra_pago__13.0	0.29	1.34	0.00	0.29	0.30	1.33	1.35	0.00	75.44	<0.005	inf
fomra_pago__DESCONOCIDO	0.68	1.98	0.01	0.67	0.70	1.95	2.01	0.00	94.58	<0.005	inf
casado__1	-0.09	0.92	0.00	-0.09	-0.08	0.91	0.92	0.00	-35.17	<0.005	897.63
edad__32-41	-0.01	0.99	0.00	-0.02	-0.01	0.98	0.99	0.00	-5.09	<0.005	21.40
edad__41-53	-0.08	0.92	0.00	-0.08	-0.07	0.92	0.93	0.00	-28.25	<0.005	580.77
edad__53 +	-0.08	0.92	0.00	-0.09	-0.08	0.92	0.93	0.00	-27.17	<0.005	537.79
edad__DESCONOCIDO	0.06	1.06	0.00	0.05	0.07	1.05	1.07	0.00	12.09	<0.005	109.39
precio__11.990-15.990	0.02	1.02	0.00	0.02	0.03	1.02	1.03	0.00	8.76	<0.005	58.85
precio__15.990 +	0.10	1.10	0.00	0.09	0.10	1.10	1.11	0.00	33.30	<0.005	805.10
precio__DESCONOCIDO	-0.76	0.47	0.28	-1.31	-0.21	0.27	0.81	0.00	-2.69	0.01	7.12
n_per__0-2	-0.04	0.96	0.00	-0.04	-0.03	0.96	0.97	0.00	-15.66	<0.005	181.27
n_per__2-3	-0.01	0.99	0.00	-0.02	-0.01	0.98	0.99	0.00	-4.90	<0.005	20.01
canal__Mayorista	0.54	1.71	0.01	0.53	0.55	1.69	1.73	0.00	95.63	<0.005	inf
canal__Tiendas Express	-0.02	0.98	0.00	-0.03	-0.01	0.97	0.99	0.00	-6.26	<0.005	31.30
canal__Tiendas Propias	-0.08	0.93	0.00	-0.09	-0.07	0.92	0.93	0.00	-16.01	<0.005	189.14
canal__Venta Remota	0.28	1.32	0.00	0.27	0.29	1.31	1.33	0.00	87.10	<0.005	inf
canal__Web	0.32	1.38	0.01	0.31	0.34	1.36	1.40	0.00	47.88	<0.005	inf
lineas_rut_6+	0.17	1.18	0.02	0.14	0.20	1.15	1.22	0.00	11.07	<0.005	92.20
lineas_rut_DESCONOCIDO	-0.34	0.71	0.01	-0.36	-0.33	0.70	0.72	0.00	-41.50	<0.005	inf

Figura 1: Resumen modelo predictivo Cox Proportional Hazard

Anexo B: Tablas comparativas

En las siguientes tablas se realizan comparaciones. En primer lugar se comparan tres metodologías altamente utilizadas en proyectos de ciencia de datos. En la segunda tabla se realiza una comparación de modelos que pueden ser implementados para el cumplimiento del proyecto.

Aspecto	CRISP-DM	SEMMA	KDD
Flexibilidad	**Alta** **Estructura Modular** Cada fase se considera como un módulo independiente con objetivos claros, lo que permite adaptar el enfoque según las necesidades específicas del proyecto	**Moderada** **Enfoque Modular** Aunque ofrece adaptabilidad, la estructura de SEMMA puede no ser tan ampliamente aplicable en diversas situaciones.	**Moderada** **Diversas Etapas** Su flexibilidad se limita a las etapas del proceso y puede no ser tan granular como el enfoque de estructura modular de CRISP-DM.
Fases Principales	1. Entendimiento del Negocio 2. Comprensión de los Datos 3. Preparación de los Datos 4. Modelado 5. Evaluación 6. Despliegue	1. Muestreo 2. Exploración 3. Modificación 4. Modelado 5. Evaluación	1. Selección de Datos 2. Preprocesamiento 3. Transformación y Minería 4. Evaluación 5. Presentación de Resultados
Iteración	Sí, en todas las fases	Sí, en fases específicas	Sí, en fases específicas

Enfoque en el Negocio	Fuerte	Moderado	Moderado
Documentación	Estructurada y detallada	Variable	Variable
Comunidad y Soporte	Amplia	Moderada	Moderada
Adopción en la Industria	Amplia	Moderada	Moderada

Tabla 1: Comparativa entre metodologías KDD, CRISP-DM y SEMMA [11]

Aspecto	Análisis de Supervivencia	Bosque Aleatorio	Redes Neuronales Artificiales
Análisis de Tiempo hasta Evento	Modelo explícito del tiempo hasta que ocurre un evento.	No se enfoca inherentemente en el análisis de tiempo hasta el evento.	No se enfoca inherentemente en el análisis de tiempo hasta el evento.
Manejo de Datos Censurados	Diseñado específicamente para manejar datos censurados.	Puede manejar datos censurados en cierta medida.	Puede manejar datos censurados, pero puede requerir cuidados adicionales.
Predicciones Dinámicas	Permite predicciones dinámicas, actualizando las curvas de supervivencia.	Típicamente requiere reentrenar el modelo completo con nuevos datos.	Puede ser re entrenada, pero el proceso puede ser computacionalmente intensivo.
Identificación de Factores de Riesgo	Identifica factores de riesgo examinando cómo las variables influyen en la tasa de riesgo.	Proporciona importancia de características, indicando la contribución de las variables.	Asigna pesos a las características, pero la interpretación puede ser compleja.
Análisis de Cohortes	Puede extenderse fácilmente para realizar análisis de cohortes.	Puede realizar análisis de cohortes agrupando datos.	Puede realizar análisis de cohortes, aunque puede ser menos directo.

Manejo de Covariables Cambiantes en el tiempo	Puede manejar covariables cambiantes en el tiempo.	Puede manejar covariables cambiantes en el tiempo, pero puede requerir consideración.	Puede manejar covariables cambiantes en el tiempo, pero puede requerir consideración.
Interpretabilidad	Proporciona curvas de supervivencia interpretables y tasas de riesgo.	Ofrece importancia de características, indicando la contribución de las variables.	Menos interpretable debido a estructuras complejas y no lineales.
Requisitos de Datos	Requiere datos de tiempo hasta el evento, incluyendo observaciones censuradas.	Adecuado para datos tabulares con características y etiquetas.	Típicamente requiere cantidades sustanciales de datos.
Complejidad Computacional	Complejidad computacional moderada.	Complejidad computacional moderada, con posibilidad de paralelización.	Complejidad computacional alta, especialmente con redes grandes.
Manejo de No Linealidades	Flexible en el manejo de relaciones no lineales.	Capaz de capturar patrones no lineales inherentemente.	Inherentemente no lineal, pero puede requerir ajustes cuidadosos.

Adecuación para Conjuntos de Datos Pequeños	Adecuado para conjuntos de datos pequeños, especialmente con censura.	Generalmente adecuado para conjuntos de datos pequeños a medianos.	Puede requerir más datos, especialmente para arquitecturas complejas.
Ensamble/Combinación con Otros Modelos	Puede combinarse con otros modelos o utilizarse en conjuntos para mejorar el rendimiento predictivo.	Comúnmente se utiliza en métodos de conjunto como bagging y boosting.	Puede ser parte de arquitecturas de conjunto, pero se necesita precaución para evitar sobreajuste.

Tabla 2: Comparación Árboles de decisión, Análisis de Supervivencia y Redes Neuronales. [1] [8]

Anexo C: Evolución de las portaciones en la industria

Los siguientes gráficos muestran la evolución y las variaciones de las portaciones desde la aplicación de la Ley de Portabilidad. En los primeros dos gráficos se muestra la evolución por empresa y, en el último, por tipo de servicio (prepago o postpago).

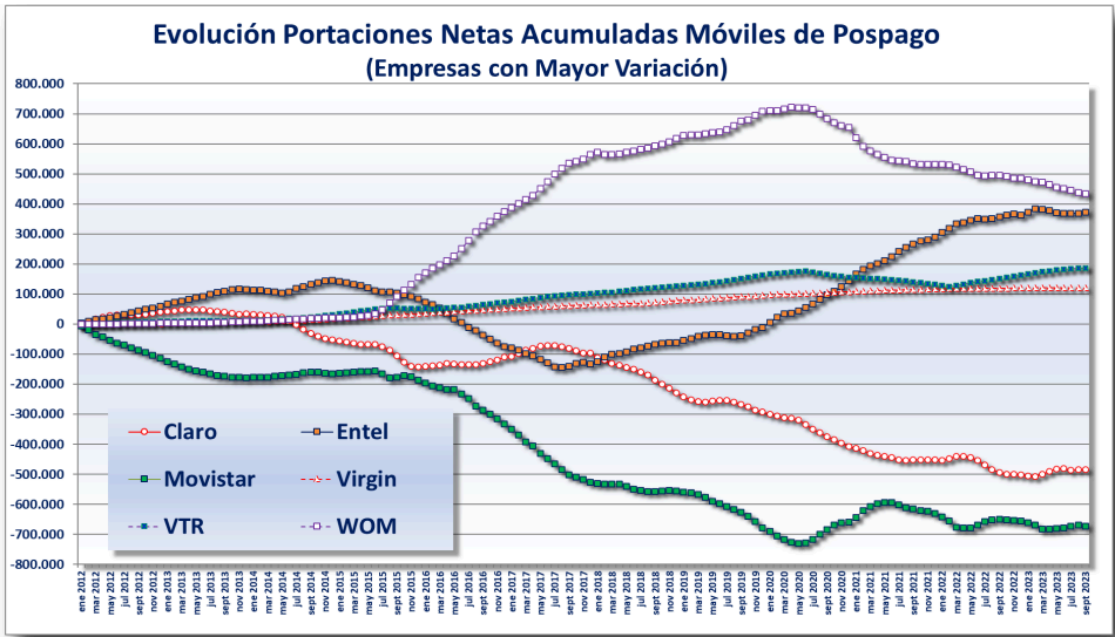

Figura 2: Evolución de las portaciones netas en el segmento de telefonía móvil postpago. Gráfico obtenido a partir del último Reporte Mensual de Portabilidad, septiembre 2023.

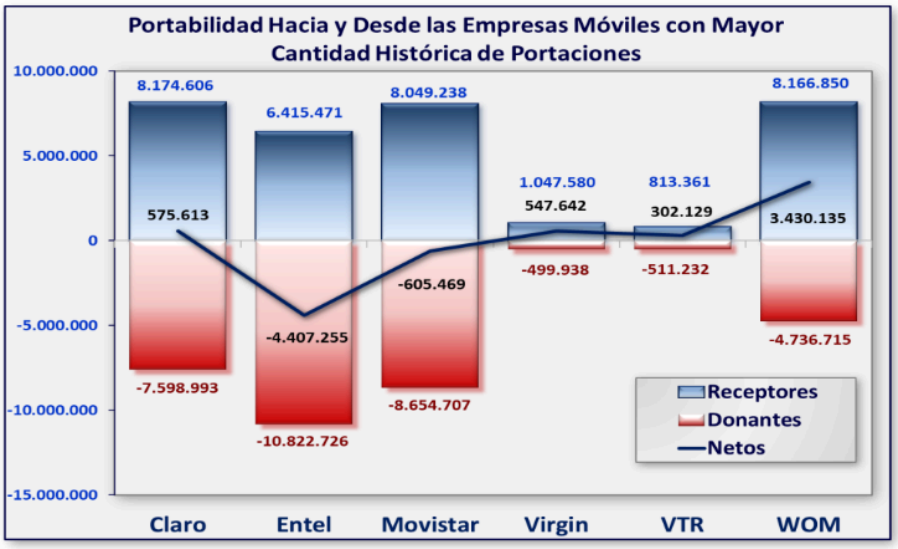

Figura 3: Evolución de las portaciones históricas en el segmento de telefonía móvil pospago. Gráfico obtenido a partir del último Reporte Mensual de Portabilidad, septiembre 2023.

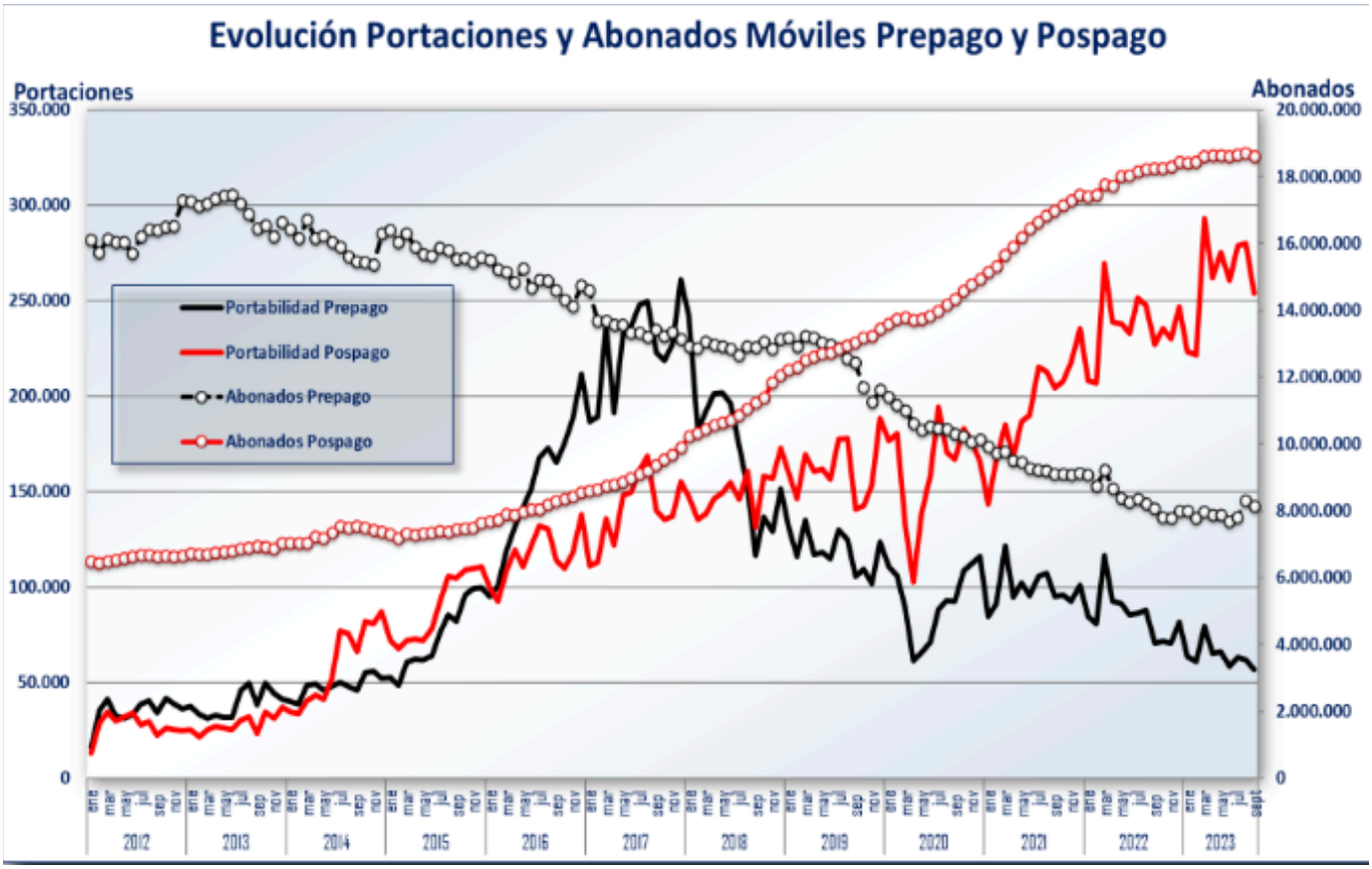

Figura 4: Evolución de las portaciones históricas para distintos segmentos. Gráfico obtenido a partir del último Reporte Mensual de Portabilidad, septiembre 2023.

Anexo D: Variables del modelo

En la siguiente tabla se muestran las variables utilizadas en el modelo, junto a una breve descripción, el origen de la variable y los distintos valores que puede tomar.

Variable	Descripción	Variable de origen	Valores
Edad del titular	Edad del titular al momento de la habilitación	Fecha_nacimiento	Valores enteros mayores a 0.
Canal	Agrupación de los valores únicos de la variable canal, en segmentos más generales	Canal	S2S, ONLINE, TIENDAS, OUTBOUND, DESCONOCIDO,OTRO,
Segmento edad	Las edades se dividen en 4 intervalos según el intervalo al que pertenezcan, los que se definieron según percentil 25, 50 y 75.	Edad_titular	5) 0-32 6) 32-41 7) 41-43 8) 53+
Trimestre	Trimestre en el que se realiza la habilitación	Periodo	T1: enero, febrero, marzo T2: abril, mayo, junio T3: julio, agosto, septiembre T4: octubre, noviembre, diciembre
gestión canal	Se agrupan los valores originales de la variable en los principales canales	ges_grupo_Canal	S2S, ONLINE, TIENDAS, OUTBOUND, DESCONOCIDO,OTRO

Tabla 3: Descripción de las variables modificadas para el modelo

Nombre Variable	Descripción	Formato	Valores
Periodo	Fecha de habilitación de la línea	número entero (yyyymm)	Se consideran las fechas desde julio de 2021 hasta la actualidad
Tipo Actividad Comercial	Describe el tipo de habilitación	cadena de caracteres (VARCHAR)	Migración prepago a postpago Port in: línea portada desde otra compañía Habilitación: Línea nueva (no existía previamente)
número móvil	Número de la línea móvil	valor entero	Valor único que identifica a cada línea móvil
rut_num	RUT asociado al titular de la línea	valor entero	valor único que representa el RUT del titular
Ciudad	Ciudad donde vive el titular	VARCHAR	Nombre de la ciudad
comuna	Comuna del titular titular	VARCHAR	Nombre de la comuna
tipo_linea	Indica si la línea está asociada a un cliente de la empresa o es un nuevo cliente	VARCHAR	Additional: corresponde a una línea adicional Base: primera línea habilitada
precio_destino	Representa el precio del plan contratado	Valor entero	Valores entre 0 y 69.990
Canal	Especifica el canal por el cual se generó la suscripción	Cadena de caracteres	Call Center, Venta Remota, Tiendas Propias, Tiendas Express, Web,

			Mayorista, Gtr, Agente Mall, Ivr
Descuento	Descuento aplicado en la suscripción	Cadena de caracteres	Tipo de descuento: A ó B
Periodo_fuga	Fecha de fuga de la línea	Valor entero	Fechas en formato YYYYMM ejemplo: 202107
total_hogar	de personas viviendo con el titular	Valor entero	número entero >= 0
Fecha_nacimiento	Fecha nacimiento del titular	Valor Fecha	Fechas formato dd-mm-YYYY
rut_conyugue	RUT del cónyuge del titular	cadena de caracteres	Identificador RUT único

Tabla 4: Descripción de variables seleccionadas para el proyecto.