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

MODELO PREDICTIVO DE RIESGO PARA PRODUCTOS DE LÍNEA DE NEGOCIO HOGAR DE UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL

DANIELA CONSTANZA BAEZA TEJO

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
NICOLÁS CISTERNAS GONZÁLEZ

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: DANIELA CONSTANZA BAEZA TEJO
FECHA: 2023
PROF. GUÍA: PABLO MARÍN VICUÑA

MODELO PREDICTIVO DE RIESGO PARA PRODUCTOS DE LÍNEA DE NEGOCIO HOGAR DE UNA EMPRESA DE TELECOMUNICACIONES

Debido a la creciente demanda de fibra óptica en Chile, una empresa de telecomunicaciones que cuenta con este servicio ha debido realizar grandes inversiones en esta tecnología con el objetivo de mantener su posición en la industria y aumentar sus ventas. Esta situación ha conllevado a una tendencia al alza de los indicadores de riesgo de este servicio, es por este motivo que mediante el siguiente trabajo se busca estimar el nivel de riesgo de los servicios de fibra óptica con el fin de implementar estrategias que permitan accionar para generar mayores ingresos y movilizar los niveles de riesgo hacia rangos deseables.

Para identificar el comportamiento de no pago se generan modelos de admisión binarios, donde la variable objetivo presenta un valor igual a 1 en caso de que un cliente adeude un saldo 60 días posterior al vencimiento de su segunda factura, y 0 en caso contrario. A partir del output de los modelos se construye la QNP, que indica el porcentaje de clientes que no pagan el servicio de una camada específica.

Para afinar la predicción del comportamiento de no pago, se segmentan los clientes en función de su antigüedad con la empresa: nuevos, semi nuevos y antiguos; los cuales representan respectivamente el 55 %, 5 % y 40 % de los casos. La metodología utilizada corresponde a CRISP-DM. Se generaron 18 escenarios de modelamientos distintos para cada segmento de clientes y dado que solo el 4 % de los datos presentaba el comportamiento estudiado se utilizan las técnicas de balanceo Over & Under sampling y Weight of Classes.

Para los clientes **nuevos** el algoritmo de clasificación binaria que entrega el mejor desempeño corresponde a un Random Forest con AUCPR igual 0,3213. La estrategia de venta para estos clientes contempla el cierre de la venta al tramo más riesgoso, aumentando las utilidades anuales en 18 millones y reduciendo la QNP2F60 en un 8,9 %.

En cuanto a los clientes **semi nuevos**, el modelo con mejor rendimiento corresponde a un Random Forest que entrega un AUCPR del 0,4303. La estrategia para este segmento consiste en el cierre de la venta a los 2 grupos más riesgosos de clientes, de esta forma es posible aumentar las utilidades anuales en 13 millones y reducir la QNP2F60 un 15 %.

Finalmente, para el segmento de clientes **antiguos** el algoritmo seleccionado corresponde a un Gradient Boosting Machine que obtiene un AUCPR igual a 0,2803. La estrategia propuesta para este perfil de clientes consiste en mantener todos los tramos de riesgo abiertos a la venta, hasta que el grupo más riesgoso alcance una QNP2F60 de 12 %.

«En el futuro, no habrá mujeres líderes. Solo habrá líderes».
— Sheryl Sandberg

Agradecimientos

Me gustaría comenzar agradeciendo a todas las mujeres que han sido un pilar fundamental en mi vida. A mi madre, Graciela, quien siempre se ha caracterizado por su fortaleza en los momentos difíciles y me ha enseñado con mucha comprensión y amor lo importante de escuchar a otros y el valor de lo justo. A la otra madre y hermana que la vida me dio, Jenny y Andrea, ustedes me han acompañado durante toda la vida e incansablemente han apoyado mis sueños y metas, gracias por abrirme las puertas de su hogar y enseñarme que el amor sincero esta por sobre las convenciones habituales de familia. A mi querida Magaly, gracias por sus cuidados y reconocer en mi un potencial, pero por sobre todo gracias por enseñarme que la disciplina es la clave de todo lo que quiero cultivar en mi vida.

Gracias a mi hermana Maraí por impregnarme de su optimista y alegre forma de ver la vida, siento un profundo respeto y admiración por la fortaleza que día a día demuestras. Gracias también a mi hermano Diego, quien siempre me acompaña en conversaciones profundas y llenas de humor un tanto negro, estoy ansiosa por descubrir el maravilloso hombre que serás.

Por supuesto a Giamnfranko, el amor que me ha apoyado inquebrantablemente en cada proyecto que he comenzado incluso desde antes de la etapa universitaria, gracias por tus consejos y palabras de aliento en los momentos difíciles, pero sobre todo por elegir ser mi compañero de tantas aventuras. A la familia del Giamn, que también hoy son mi familia, gracias por la calidez de su hogar y el cariño tan sincero que me han entregado todos estos años.

Por supuesto gracias a mis amadas amigas y amigos, quienes han jugado un rol muy importante en este proceso de cierre. En ustedes he encontrado tranquilidad y risas, también comprensión y ayuda. Siempre estaré muy agradecida de tener la posibilidad de observar como siguen floreciendo y conviertiendose en seres humanos maravillosos.

Finalmente, quiero agradecer al gran equipo de trabajo que me brindó la posibilidad de realizar esta memoria; pero por sobre todo me gustaría agradecer a Constanza, Natalia y Andrea, por su guía y compañía en mi desarrollo profesional. Ha sido un honor poder trabajar con mujeres que me inspiran cada día con su profesionalismo y calidez.

A cada una y uno de ustedes, gracias por decidir ser parte del bosque de mi vida.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Caracterización de la empresa

El Trabajo de Título está enmarcado dentro una destacada compañía de telecomunicaciones nacional. Esta cuenta con presencia en Chile desde 1964 y en Perú desde el año 2014 [1]; se perfila como una corporación **líder en tecnología y telecomunicaciones** abarcando a más 20,1 millones de clientes sumando los dos países [2] durante el 2021.

De acuerdo a la memoria 2021 [2] de la compañía, los ingresos de esta alcanzaron los $ 2.460.119 millones de pesos chilenos, además aumenta su cantidad de clientes un 12,4 % con respecto al año 2020; para lograr esto cuentan con 12.248 trabajadores en conjunto con una fuerte inversión en nuevas tecnologías como el 5G o fibra óptica, alcanzando en esta última categoría más de 6000 km desplegados a través de Chile.

La empresa declara [3] que su **propósito** consiste en: ***acercar a la sociedad y a sus clientes las infinitas posibilidades que da la tecnología***.

Los productos y servicios que se ofrecen dependen del segmento de clientes, en Chile la empresa cuenta con 5 tipos de perfiles que se satisfacen mediante los siguientes mercados: **Personas**, **Empresas**, **Corporaciones**, **Mayoristas** y **Call Center**, presentados respectivamente de acuerdo a la cantidad de ingresos que aportan a la empresa.

El trabajo se enmarca dentro del mercado[1] **personas**, en este se encuentran todos los productos (*véase* Figura 1.1) dedicados a satisfacer las necesidades de **personas naturales**.

[1]El desglose de cada línea de negocio con sus respectivos productos y servicios se encuentran en las páginas 36-39 de la Memoria Anual Integrada 2021 (link disponible en Bibliografía).

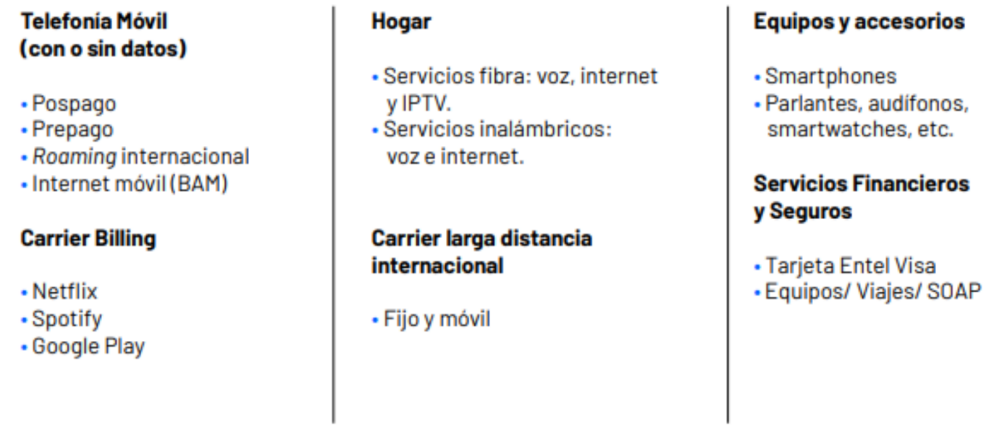


Figura 1.1: Productos ofrecidos para el mercado personas de la compañía. Extraído desde Memoria Anual 2021 [2].

Por otro lado, tal como se indica en un inicio, el perfil de la empresa consiste en una compañía líder en tecnología y telecomunicaciones, esta situación se ha visto reflejada en la valorización de las acciones de la empresa donde ICR en su informe de Marzo del 2022[4] destaca su **infraestructura de red como una ventaja competitiva** en su sector, ya que esta le ha permitido posicionarse como una de las empresas con mayor cobertura entregando una alta calidad en sus servicios.

1.2. Sector industrial

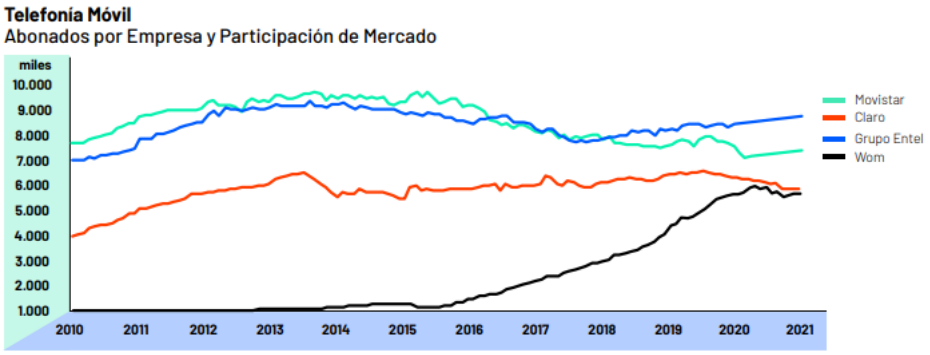
El sector industrial en el cual participa corresponde a **la industria de las tecnologías y las comunicaciones**, en Chile esta industria se posiciona como una de las más relevantes sobrepasando los 59 millones de contratos activos durante el 2021 [2]. Este rol central se potenció debido a los nuevos escenarios de teletrabajo conducidos por la pandemia COVID-19, situación que se ve reflejada en un aumento del 30,7 % del tráfico de datos móviles durante el año 2021 [2]; en cuanto al sector hogar, durante el 2021 se estima que el 67,9 % de los hogares en nuestro país contaba con acceso a internet fijo en contraste al 48 % que se alcanzó durante el 2017 [2].

La constante necesidad de los usuarios por un servicio cada vez más potente y con menor porcentaje de interferencias, ha llevado a los participantes de este mercado a realizar grandes inversiones tecnológicas acorde a las nuevas demandas de los clientes; también es relevante destacar que la industria cuenta con importantes ingresos, registrando a septiembre del 2021 una entrada de 7,9 miles de millones de dólares, monto equivalente al 1,7 % del PIB chileno [2].

En la Figura 1.2 se puede observar los productos más relevantes en la industria con sus respectivos porcentajes de participación. Además, dentro de los principales actores de la industria se encuentra: **Movistar, Claro, Wom, VTR, Virgin, Mundo Pacífico y Grupo Entel**.

Figura 1.2: Porcentaje de ingresos divididos en las diferentes categorías de producto del sector industrial en cuestión. Memoria integrada 2021 [2].

Dentro del mercado de personas la compañía cuenta con una consolidada participación en los productos asociados a la telefonía móvil, posicionándose como el líder de esta categoría (*véase* Figura 1.3).



Participación de Mercado	Dic 20	Dic 21
Movistar	24,8%	24,9%
Grupo Entel	31,9%	32,2%
Claro	20,8%	21,3%
Virgin	0,6%	0,4%
WOM	20,8%	20,0%
VTR	1,1%	0,9%
Otros	0,1%	0,3%

Figura 1.3: Principales participantes del sector industrial con sus respectivos porcentajes de penetración durante 2021 para productos móviles. Memoria integrada 2021 [2].

A su vez, en lo que respecta a productos hogar la empresa no se encuentra dentro de los primeros lugares, en Chile este segmento de productos es liderado por VTR [2]. Sin embargo, es relevante destacar que a diciembre del 2021, la compañía aumentó en un 12,7 % sus conexiones fijas lo que se tradujo en una penetración de 21,3 por cada 100 habitantes del territorio nacional [2]. Del total de conexiones, el 89,6 % correspondió a accesos residenciales y el porcentaje restante a conexiones empresariales.

Finalmente, es importante destacar que el marco regulador de las acciones de las empresas de telecomunicaciones está definido por la Subsecretaría de Telecomunicaciones (SUBTEL), esta tiene por objetivo: *coordinar, promover, fomentar y desarrollar las telecomunicaciones en Chile, transformando a este sector en motor para el desarrollo económico y social del país* [5]. Dentro de las iniciativas más recientes que lidera esta subsecretaría se encuentra la instalación del 5G como un medio de democratización para el acceso al internet en nuestro país [2].

[2]En la página oficial de SUBTEL es posible encontrar un resumen con las políticas públicas más relevantes. Revise el siguiente enlace para más información: https://www.subtel.gob.cl/normativa-tecnica-internet/.

1.3. Desempeño organizacional

De acuerdo al informe de ICR Chile [4], la compañía se encuentra en un **estado de consolidación**; sin embargo, dado que la industria es altamente competitiva, la empresa se ha reinventado con el objetivo de mantener su posición en el mercado.

En el informe de Marzo de 2022 [4], se indica que la empresa ha mantenido una trayectoria del market share de productos de voz móviles en torno al 32 %, mientras que para los productos de datos móviles este porcentaje alcanza el 35 %.

Además, gracias a la estrategia de negocio que ha permitido apalancar grandes inversiones en fibra óptica, la compañía ha crecido en la penetración del mercado hogar pasando de un **1,7 % en el 2017 al 6,5 % durante el 2021**; dado que la sociedad ganó parte de las licitaciones de las bandas 5G se espera una comportamiento estable, que permita mejorar la posición en la que se encuentra actualmente en este mercado.

Por supuesto, esta etapa de madurez se caracteriza por la cantidad estable de los ingresos (*véase* Figura 1.4), la cual ha significado un margen de EBITDA cercano al 30 % en los últimos 3 ejercicios anuales.

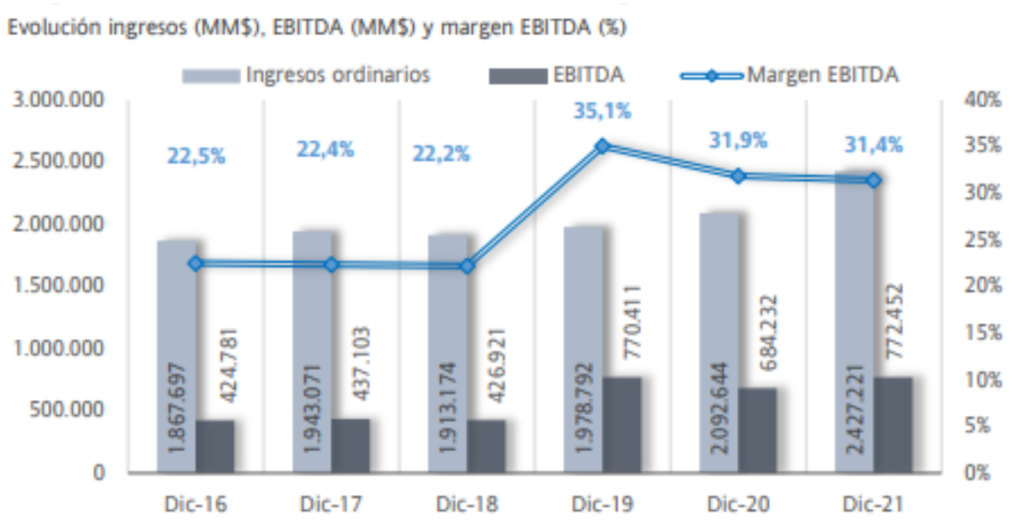


Figura 1.4: Ingresos orgánicos, EBITDA y margen EBITDA desde dicimenbre 2016 a diciembre del 2021. Elaborado por ICR Chile (2022)[4].

Capítulo 2

Problemática

2.1. Áreas involucradas

Transversal a diferentes líneas de negocio de la empresa se ha construido una comunidad denominada "Analytics", la cual cuenta con diferentes perfiles profesionales a cargo del estudio, modelamiento y reportería de la información de sus mercados objetivos.

En función del mercado al cual pertenecen sus clientes existen tres comunidades de excelencia operacional (CoE) (*véase* Figura 2.1).

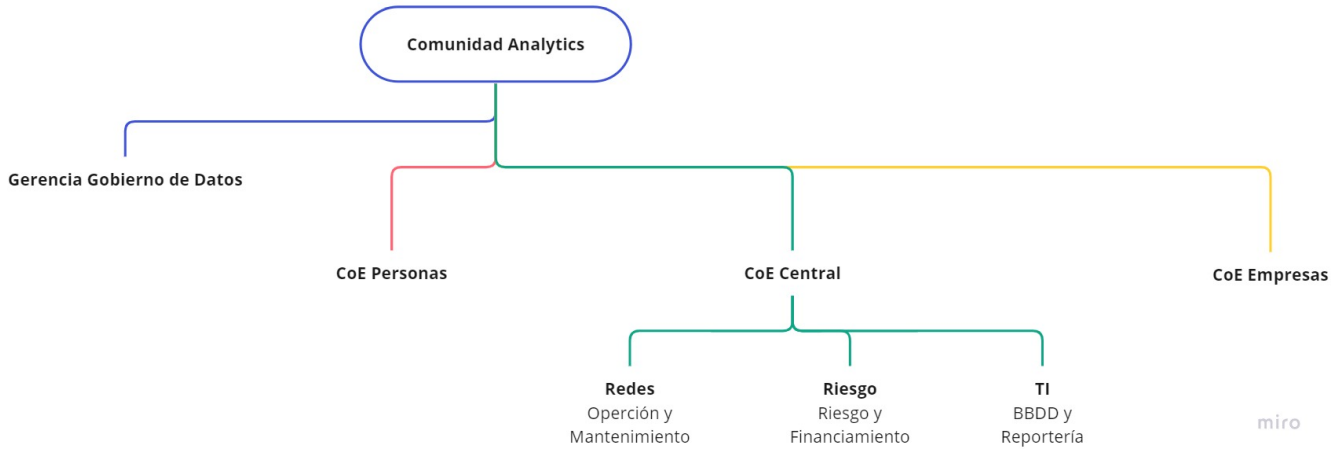


Figura 2.1: Estructura comunidad Analytics. Elaboración propia.

El CoE Personas trabaja realizando análisis y modelos para mercado personas, mientras que CoE Empresas realiza las mismas acciones para el mercado con el mismo nombre.

El trabajo de título esta inserto en el CoE Central, este último se encuentra al interior de la Vicepresidencia de Tecnología y Operaciones. La unidad está compuesta por 3 equipos (*véase* Figura 2.1), los cuales persiguen los siguientes objetivos:

1. Implementar las mejoras de seguridad y buenas prácticas para toda la comunidad Analytics.

2. Realizar modelos de predicción de riesgo y financiamiento para la toma de decisiones estratégicas.

3. Realizar análisis de data y construcción de modelos para mejorar la perfomance del soporte técnico de la compañía.

Es importante destacar que las labores a realizar se enmarcan dentro de los proyectos enfocados en la predicción del riesgo. El CoE Central cuenta con 13 profesionales, de los cuales 5 están dedicados a los proyectos de este tipo.

Los clientes para el CoE Central son en esencia otras áreas de la compañía, las cuales dependen del proyecto que se este realizando. En el caso de esta iniciativa existen dos actores relevantes, en primer lugar se encuentra el **Área de Riesgo**, la cual es responsable de **vender al menor riesgo posible**, para lograr esto debe delimitar las reglas de negocios de los diferentes mercados y mitigar comportamientos riesgosos de productos y servicios.

En un rol secundario se encuentra al **Área Comercial de productos Hogar**, este equipo esta a cargo de **impulsar la venta de productos hogar**; además, de trabajar en la fidelización de los actuales clientes de la línea de negocio.

Tango Riesgo como el área Comercial de hogar están constantemente en comunicación, ya que las políticas que un equipo determine pueden impactar significativamente en los resultados esperados del otro equipo; por lo que es de suma importancia que la estrategias definidas surjan a partir de un acuerdo entre ambas partes.

2.2. Definiciones claves

Antes de profundizar en la comprensión del problema es relevante clarificar las siguientes definiciones:

- **Línea hogar:** referencia a la línea de negocio hogar de la compañía. Esto incluye los servicios fibra (voz, internet y TV full [1]), servicios inalámbricos (voz e internet) y TV light [2].

  Estos servicios son vendidos en paquetes, también denominados **bundles**. Cuenta con planes **mono** los cuales solamente son un producto, mientras que los **dúos** y **tríos** cuentan con dos 2 o 3 productos respectivamente.

- **Camada:** corresponde a una ventana de tiempo mensual; todos los clientes de una camada corresponderán a los clientes que adquieren un producto durante un mes específico[3].
- **Saldo:** se entenderá por saldo la deuda (CLP) que tiene un cliente con la compañía.
- **Default:** variable objetivo, tomará un valor igual a 1 si el cliente presenta un saldo mayor a cero, y cero si es que no presenta un saldo adeudado.
- **Quantity No Pago (QNP_{pi}):** Suma de clientes que presentan un default igual a 1 para un determinado servicio p y camada i, por sobre el total de clientes que accedieron al mismo servicio p durante la misma camada i.

$$QNP_{pi} = \frac{\text{Cantidad de clientes que no pagan}_{pi}}{\text{Total de clientes de la camada}_{pi}} \tag{2.1}$$

- **Tasa No Pago (TNP_{pi}):** la tasa de No Pago para un determinado servicio p y camada i, es calculada como: el monto adeudado para un determinado servicio p y camada i, por sobre los ingresos del mismo servicio en la misma camada.

$$TNP_{pi} = \frac{\text{Saldo camada}_{pi}}{\text{Ingresos camada}_{pi}} \tag{2.2}$$

- **Clientes nuevos** (CN_{pi})**:** son todas aquellas personas de una camada i que al momento de contratar un servicio p sean personas completamente nuevas para la compañía.
- **Cliente semi nuevos**(CSN_{pi})**:** corresponde a cualquier persona de una camada i, que al momento de contratar el servicio p poseía una antigüedad entre 1 y 5 meses con la empresa.
- **Clientes antiguos** (CV_{pi})**:** son todas aquellas personas naturales de una camada i, que al momento de contratar un servicio p de la línea sean consumidores de otro producto de la compañía por al menos seis meses.

[1] Entiéndase por TV full el servicio de televisión que utiliza la red fibra dispuesta en un domicilio, incluye la programación completa que dispone la empresa.

[2] Entiéndase por TV light un servicio de televisión streamming similar a plataformas como Netflix.

[3] Esto implica que un mismo cliente puede pertenecer a diferentes camadas, ya que la persona puede acceder a diferentes productos o servicios.

2.3. Problemática

El riesgo de los diferentes servicios puede verse de manera global, considerando tanto a clientes nuevos como antiguos, y todos los productos de la línea. Pero usualmente se estudia desglosando las Tasas de No Pago (TNP) por tipo de servicio (fibra, inalámbrico, TV light) y/o antigüedad del cliente (nuevos, semi nuevos o antiguos).

A mediados del 2020 Riesgo observa aumentos en las Tasa de No Pago (TNP) de la línea hogar, siendo los servicios inalámbricos y TV light las categorías más preocupantes (*véase* Figuras 2.2 y 2.3). En un inicio este efecto fue asociado a las limitaciones producidas por la pandemia; sin embargo, con el transcurso del tiempo no se ha observado una normalización de las tasas de no pago.

Figura 2.2: Evolución TNP 30 días primera factura para el periodo estudiado, desagregado por servicio.

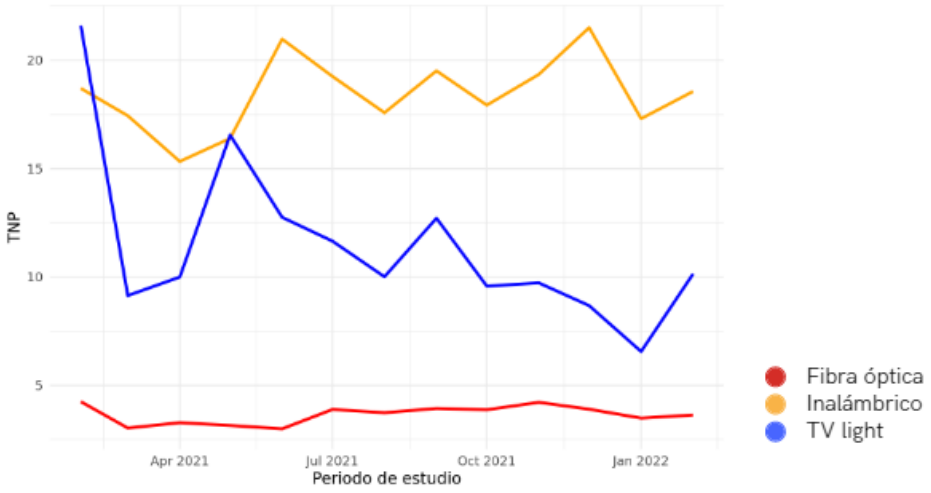
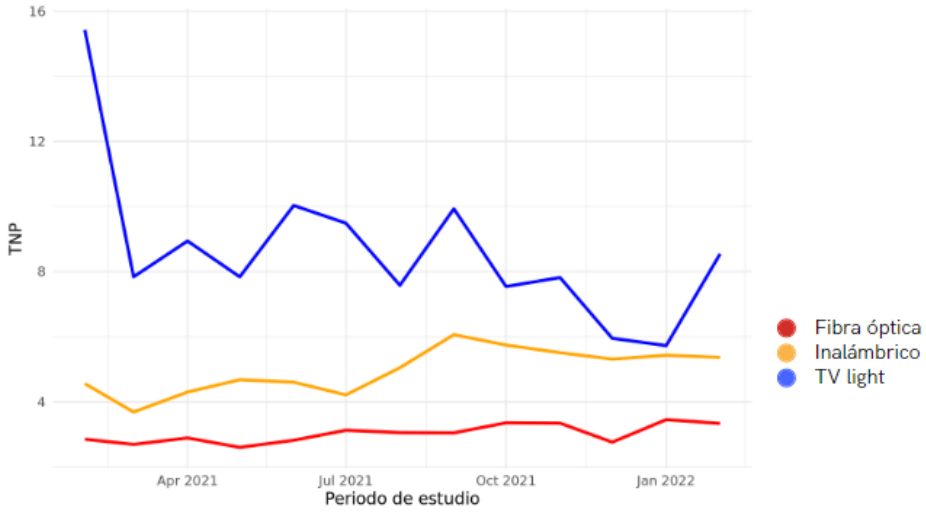


Figura 2.3: Evolución TNP 60 días segunda factura para el periodo estudiado, desagregado por servicio.



Para la exploración inicial de este problema se trabaja con las camadas comprendidas entre febrero 2021 a febrero 2022, correspondiente a 223.625 registros. Además, en lo que respecta a la evolución de las tasas de no pago se observa una diferencia en el comportamiento de las TNP a 30 y 60 días, -esto es sin diferenciar entre la antigüedad del cliente-.

Para la tasa a 30 días posterior a la primera factura, el servicio inalámbrico presenta el comportamiento más riesgoso ya que sus TNP son las más altas (*véase* Figura 2.2); sin embargo, si se observa el comportamiento a 60 días vemos que inalámbrico comienza a tener un comportamiento bastante similar a fibra (*véase* Figura 2.3), se cree que este comportamiento se debe a que los clientes de productos inalámbricos presentan algunos días de mora en el pago de sus boletas, pero poseen interés en mantener activo el servicio por lo que prontamente regularizan su situación.

Por otro lado TV light mantiene un comportamiento bastante riesgo, con valores entre 6 % y 20 % (*véase* Figuras 2.2 y 2.3). A partir de esto es posible indicar que esta categoría efectivamente presenta el comportamiento más preocupante, ya que los clientes de la línea no están interesados en regularizar los servicios; por lo que, la mala identificación entre buenos y malos pagadores toma aún más relevancia.

Finalmente, se tiene el comportamiento de los servicios asociados a fibra óptica, los cuales tanto para las tasas de no pago a 30 y 60 días presentan un comportamiento más estable (*véase* Figuras 2.2 y 2.3). Al igual que en los servicios inalámbricos, se observa una disminución en el valor promedio de la tasa a 60 días con respecto a la tasa de 30 días; se espera que este comportamiento responda a la misma lógica planteada para los productos inalámbricos.

Debido a que la cobertura de fibra óptica durante el periodo de exploración inicial se concentra en sectores geográficos donde habitan personas con mayores ingresos, tiene sentido que sus tasas de pago sean menores. Esta situación, comenzará a cambiar drásticamente durante el 2023, ya que las inversiones realizadas en esta tecnología los años previos permitirán una expansión importante de la cobertura de este servicio durante el próximo año. De la mano con el último punto, todos los esfuerzos estratégicos del Área Comercial Hogar estarán situados en fibra ya que se espera que eventualmente este sea el único servicio de la línea.

Al profundizar, se logra identificar que el problema que generaba estos efectos se debe a una **discriminación ineficaz entre los buenos y malos pagadores de la línea de negocio hogar** (*véase* Anexo A: Árbol de problemas).

También se detectan algunos efectos secundarios como la acumulación de pérdidas debido a que se disponen recursos y personal para la instalación de servicios que finalmente no son cancelados; además, de que al no discriminar correctamente a los buenos pagadores se pierde la posibilidad de brindar un mejor servicio a aquellos clientes de la cartera de productos que poseen un buen comportamiento de pago.

Por otro lado, dentro de las causas raíces del problema, Analytics es capaz de hacerse cargo de volver más eficiente el pronóstico de comportamiento de pago de los clientes de la línea de negocio. Este sistema no ha sido actualizado desde el 2018, por lo que no contempla la instauración de los servicios fibra, los cuales comenzaron a venderse a finales del 2019; tampoco es capaz de estratificar el riesgo de los clientes, con el objetivo de afinar la estrategia de venta a utilizar.

Capítulo 3

Objetivos

3.1. Objetivo general

Evaluar el nivel de riesgo de los servicios fibra óptica de una empresa de telecomunicaciones, mediante la predicción del comportamiento de no pago de los clientes de la cartera de servicios.

3.2. Objetivos específicos

Para lograr el objetivo general, se definen los siguientes objetivos específicos:

1. Identificar variables relevantes para la predicción del no pago de clientes.

2. Evaluar el resultado de los diferentes modelos construidos para seleccionar el que entregue el mejor desempeño.

3. A partir del resultado del modelo seleccionado, proponer una estrategia de ventas que permita mantener el riesgo de los productos fibra en niveles deseados.

4. Elaborar un plan de implementación que identifique los riesgos de la ejecución del modelo, con el fin de garantizar una automatización exitosa de este.

Capítulo 4

Alcances del trabajo

- Los modelos a implementar solo cuentan con datos del mercado hogar de Chile, lo que implica que las variables serán representativas solo para clientes de este mercado. Esto conlleva a que los resultados no deben ser extendidos a otro tipo de línea de negocio o cliente; por ejemplo un cliente móvil o empresas tendrá un comportamiento distinto a un cliente hogar, por lo que el modelo resultante no debe ser utilizado para estudiar el comportamiento de otras líneas de negocio.

- Los registros utilizados contienen información desde Agosto del 2020 a Mayo del 2022[1], -se escoge este mes para finalizar la captura de datos ya que al momento de modelar este contiene la última TNP 60 días posterior a la segunda factura disponible-.

- La selección de datos para entrenamiento y testeo es diferente a la utilizada habitualmente debido a que se eligen determinadas camadas para cada uno de los sets de datos; el motivo de esta decisión consiste en que se busca capturar los comportamiento mensuales, por lo que las camadas son lo más representativo a esta situación. Dicho esto, se usa para **entrenamiento** las camadas comprendidas entre Agosto del 2020 y Diciembre del 2021, abarcando cerca del 75 % de los datos; mientras que para la **evaluación** se emplean las camadas contenidas entre Enero a Mayo del 2022, las cuales representan el 25 % restante de los registros.

  De la mano con el punto previo, es importante destacar que dentro del negocio se asume que las diferentes camadas son independientes entre sí, por lo que se espera que el comportamiento de cada una de estas no interfiera en la conducta de otra.

- Pese a que se cuenta con información de los diferentes servicios de la línea, el foco de los modelos se encuentra en fibra. Debido a decisiones estratégicas la empresa esta buscando reducir a cero las ventas de los otros servicios; por lo que espera que a futuro solo se cuente con clientes de fibra en hogar, dejando obsoletos los modelos para los otros servicios.

- No es posible asumir causalidad entre las variables que mejor expliquen el comportamiento

[1]Se cuenta con registros previos a esta ventana de tiempo, pero no son considerados debido a que son camadas con registros considerablemente menores a las seleccionadas.

de no pago y la variable objetivo debido a que los resultados no responden a un experimento. Esto también conlleva a que los resultados presenten heterogeneidad.

Debido a que se cuenta con tiempo limitado, este trabajo no construirá un experimento para validar la existencia de causalidad de las variables, ni tampoco la ruta de trabajo para fabricar esa iniciativa.

- Este trabajo incluye las recomendaciones para una nueva estrategia de ventas y un plan de implementación, pero no contempla llevar a cabo estos puntos; esto queda a criterio de la empresa.

Capítulo 5

Marco conceptual

5.1. Selección de variables

El primer desafío que se debe enfrentar consiste en la gran cantidad de variables con las que se dispone para trabajar (más de 1300), para afrontar esta situación se investigaron acerca de diferentes métodos para la selección de variables que permitiesen acortar los tiempos de procesamiento.

- **Completitud de variables:** el primer filtro utilizado consiste en remover todas aquellas columnas que contengan más de un determinado porcentaje de missing values, en este caso la cota inferior corresponde al 80 %.

- **Valor de la información (Information Value - IV):** corresponde a un concepto ampliamente utilizado para clasificar variables relevantes para un modelo. A través de este indicador es posible determinar si una variable es relevante para el objeto de estudio [6].

$$IV = \sum \text{Distribución Buena}_i - \text{Distribución Mala}_i \cdot ln(\frac{\text{Distribución Buena}_i}{\text{Distribución Mala}_i}) \tag{5.1}$$

  1. Sea la distribución buena la cantidad de casos positivos para un grupo determinados por sobre el total de casos positivos de la muestra.
  2. De manera análoga la distribución mala de un determinado grupo corresponderá a la cantidad de casos negativos por sobre el total de casos negativos de la muestra.
  3. Se determinará un IV para cada agrupación, los cuales serán sumados para obtener el IF de la variable estudiada.
  4. El criterio teórico para mantener o eliminar una determinada variable se presenta en la Tabla 5.1, para este trabajo se utilizarán todas aquellas variables con IV mayor a 0.1.

- **Peso de la evidencia (Weight of Evidence - WoE):** es una transformación que es utilizada cuando la variable a predecir es binaria y se requiere que alguna covariable de interés sea estratificada [6].

Tabla 5.1: Criterios tentativos para la mantención y/o remoción de variables mediante Information Value. Elaborada a partir de la información propuesta por OCB [7].

Information Value	Poder predictivo
>0.5	Sospechoso
0.3 - 0.5	Predictor fuerte
0.1 - 0.3	Predictor medio
0.02 - 0.1	Predictor débil
<0.02	Irrelevante para la predicción

1. Se define la variable a estudiar ("default") y alguna covariable a estratificar (por ejemplo: "score", "edad", "ingresos").
2. Se agrupan los valores de la(s) covariable(s) de interés, de manera que las diferentes uniones tengan QNP similares. De esta forma se exige que el agrupamiento releje un comportamiento similar entre las uniones creadas.
3. Se calculan los casos positivos y negativos dentro de cada grupo de una variable.
   (a) Suponiendo que la variable x es categorizada en los grupos: G1, G2 y G3. Nos situamos en primer lugar en G1, la variable default/estudiada puede tener un total de 100 casos positivos (todos los casos positivos de la muestra), pero solo 20 de ellos estan en G1; por lo que **esto entrega un bins de positivos para el G1 de 20 %**.
   (b) El procedimiento es el mismo para determinar los casos negativos, se calcula para cada grupo de la variable x los casos que cumplan la condición negativa (default = 0) y se dividen del total de casos negativos de toda la data. Si en toda la data se tiene 50 casos negativos, y en el G1 solo 5 cumplen ese requisito, se tiene que: **el bin de casos negativos para el G1 es de un 10 %**.
4. Luego se calcula el WOE dividiendo por porcentajes de bin de casos positivos por sobre el bin de casos negativos para cada agrupación realizada de la variable x.

Es relevante indicar que tanto WOE como IV se indefinen cuando en la agrupación solo cae un tipo de caso (positivo o negativo), por lo que se impone que siempre habrán representantes de las categorías buenas y malas en cada agrupación realizada. Un canon razonables es que al menos hayan 10 elementos de cada categoría en cada uno de los diferentes grupos [7].

5.2. Modelos

Antes de profundizar en las alternativas disponibles es necesario indicar que la variable objetivo corresponderá a una **variable binaria**. Si esta toma un valor igual a 1 indica que el cliente no paga la deuda, mientras que si su valor es igual a 0 esta indicando que el cliente si paga el servicio a tiempo.

Este tipo de modelos, son conocidos como modelos de elección binaria o de clasificación. La probabilidad de ocurrencia queda explicada mediante la relación 5.2:

$$p_i = Prob(y_i = 1|x_i), i = 1, 2, ..., n \tag{5.2}$$

Dado que y solo tomará valores entre 0 y 1, se tiene que:

$$Prob(y_i = 1|x_i) = p_i \longrightarrow Prob(y_i = 0|x_i) = 1 - p_i \tag{5.3}$$

Este escenario corresponderá a una distribución Bernoulli, por lo que la esperanza condicional de y_i se presenta en la relación 5.4, mientras que la varianza condicionada de la variable se enuncia en la relación 5.5.

$$E(y_i|x_i) = 1p_i + 0(1 - p_i) = p_i \tag{5.4}$$

$$V(y_i|x_i) = p_i(1 - p_i) \tag{5.5}$$

La gama de modelos posibles a implementar es amplia, dentro de las características más relevantes a considerar para seleccionar se encuentra: la capacidad la lidear con el sobreajuste de la muestra, facilidad de interpretación y por supuesto el tiempo de implementación de cada uno de ellos. Es relevante destacar que dadas las características del trabajo a realizar, se concentran los esfuerzos por construir únicamente modelos supervisados.

5.2.1. Regresión logística

En primer lugar se tiene a los clásicos modelos lineales generalizados (GLM), el fuerte de estos radica en la simplicidad de su implementación e interpretación de resultados; sin embargo, dentro de sus debilidades encontramos la gran cantidad de requisitos que deben cumplir los datos para hacer válidos los resultados de la regresión.

Se utiliza la regresión logística debido a que esta entrega una respuesta categórica de dos niveles [8], por lo que cumple con las definiciones del problema abordado. En la ecuación 5.6 se presenta la relación que cumple esta regresión.

$$L_i = ln(\frac{p_i}{1-p_i}) = z_i = \beta_0 + \beta_i x_i \tag{5.6}$$

5.2.2. Modelos de aprendizaje automático

También se cuenta con otras herramientas centradas en el entrenamiento de los modelos [1], de acuerdo a la bibliografía recopilada se considera que las aquellas más adecuadas para este trabajo son:

- **Random Forest (RF):** formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping [9].

  Este tipo de árbol generalmente entrega buenos desempeños, sin necesidad de complejizar el modelo. A su vez, las desventajas de este modelo radican en el costo de procesamiento, ya que puede requerir una gran cantidad de tiempo [10].

- **Gradient Boosting (GB):** modelo conformado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial de forma que cada nuevo árbol pretende mejorar los errores de los árboles anteriores [9]. Posee diferentes variaciones como: XGBoost, LightGBM and CatBoost [11], siendo este último el más recomendado para variables categóricas [12].

  La principal ventaja de este algoritmo radica en su aprendizaje secuencial, ya que va generando una tasa de aprendizaje a partir de los errores de los árboles previos, lo que conlleva a un mejor resultado. Dentro de las desventajas nuevamente se encuentra el costo en tiempo ya que al igual que RF requiere de una grande tiempo considerable para el procesamiento.

- **Support Vector Machine (SVM):** metodología creada originalmente para la clasificación binaria de variables, su construcción está basada en la definición del máximo margen de clasificación.

  Para solucionar este problema, se selecciona al *maximal margin hyperplane* que corresponde a la recta más alejada de las observaciones de entrenamiento Para su obtención, se debe calcular la distancia perpendicular de cada observación a un determinado hiperplano; la menor de estas distancias (conocida como margen) determina cuan alejado está el hiperplano de las observaciones de entrenamiento [13].

  Pese a ser una de las alternativas que mejor respuesta tienen para los modelos de clasificación binaria, su gran barrera radica en la comprensión de sus resultados; además, este método requiere conocimientos más avanzados de álgebra lineal por lo que la implementación de este puede generar rechazo debido a la dificultad para comprender el resultado que el modelo genera; y es por este último motivo que no se trabaja con este algoritmo en este trabajo.

[1] Es relevante destacar que gran cantidad de las herramientas presentadas en esta sección tienen una desventaja en común: estos a diferencia de los GLM suelen ser tener una interpretación más engorrosa para los clientes del negocio, lo que puede significar una barrera en su implementación.

5.3. Balanceo de la muestra

En la sección anterior se indica brevemente que uno de los problemas a los cuales se ven enfrentados los modelos consiste en el desbalance de la data; en este caso la cantidad de personas que presentan un default igual a 1 bordea el 4 %. Para solucionar esta situación existen diferentes técnicas en las cuales se profundizará a continuación:

- **Over-sampling:** consiste en la replicación de escenarios minoritarios de la muestra, estas inclusiones son aleatorias. La desventaja de esta alternativa consiste en el sobre muestreo de los datos, llevando al modelo a ajustarse demasiado a los datos de entrenamiento lo que podría provocar que este no sea capaz de generalizar para comportamientos más específicos de la muestra [14].

- **Under-sampling:** consiste en la eliminación aleatoria de registros de la clase mayoritaria [14]. Una de las desventajas de esta herramienta radica en la pérdida de datos potencialmente útiles para el modelo.

- **Weight of classes:** esta técnica de balanceo consiste en la asignación de pesos para la clase mayoritaria y minoritaria, con el objetivo de que al momento de modelar mediante los pesos se logre identificar más certeramente la clase minoritaria.

5.4. Métricas

Para la evaluación de los resultados entregados por un modelo existen diferentes metodologías y métricas, sin embargo, en esta sección se listan las más frecuentes.

5.4.1. Matriz de confusión

La matriz de confusión (*véase* Tabla 5.2) es una estructuración que permite clasificar los diferentes resultados de un modelo, mediante esta se compara la predicción realizada versus el valor efectivo de la observación; de esta manera es posible determinar que tan buen desempeño posee el modelo.

Tabla 5.2: Configuración de matriz de confusión. Elaboración propia

Observación | Predicción	**Positivo**	**Negativo**
Positivo	VP	FN
Negativo	FP	VN

Profundicemos en las cuatro resultados posibles para los modelos que se construirán [15]

- **Verdadero Positivo (TP):** modelo predice no pago y efectivamente el cliente no paga el servicio.
- **Falso Positivo (FP):** modelo predice no pago, pero cliente si paga el servicio.
- **Falso Negativo (FN):** modelo predice pago pero cliente no paga el servicio.
- **Verdadero Negativo (TN):** modelo predice pago y cliente paga.

5.4.2. Métricas de desempeño de un modelo

A partir de la matriz de confusión se proceden a calcular las diferentes tasas: TP, FP, FN y TN; es mediante estas que construyen métricas de desempeño [15].

La **precisión** se define en la relación 5.7, es una métrica que indica la exactitud del modelo; para esto calcula el porcentaje de los que el modelo predijo como no pago y que realmente resultaron no pagar.

$$Precission = \frac{TP}{TP + FP} \tag{5.7}$$

La **sensibilidad** se representa mediante la relación 5.8, mediante esta se indica la exhaustividad del modelo; se calcula determinando el porcentaje de casos positivos (no pago) que fueron correctamente clasificados.

$$Recall = \frac{TP}{TP + FN} \tag{5.8}$$

La **exactitud** del modelo para predecir se determina mediante la relación 5.9, y corresponde al porcentaje de casos que clasificaron correctamente sobre el total.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.9}$$

La **especificidad** se calcula mediante la relación 5.10, y corresponde al porcentaje de casos que el modelo predice que si pagan y en efecto los clientes lo hacen.

$$Especificity = \frac{VN}{VN + FP} \tag{5.10}$$

A partir de las métricas previas obtendremos nociones iniciales de los resultados del modelo, identificando por ejemplo si hay una dominancia de una clase, lo que podría estar indicando un desbalance en la muestra.

5.4.3. Métrica para comparar desempeños entre modelos

Curva ROC - AUC

Para comparar los resultados de diferentes modelos se utilizará la métrica ROC (Receiver Operating Characteristics), esta métrica de error mide la capacidad del modelo para clasificar correctamente la clase mayoritaria en distintos umbrales [16], determinando el porcentaje de verdaderos negativos sobre falsos positivos para diferentes clasificaciones.

Una vez compuesta la ROC, se procede a integrar el área bajo la curva (AUC); por lo que el método suele llamarse indiscriminadamente como ROC, AUC, ROC-AUC e inclusive AUROC. Se busca que el área bajo la curva construida sea lo más grande posible: un modelo perfecto tendría un AUC igual a 1, uno pésimo tendría un AUC igual a 0, mientras que un AUC igual a 0.5 indica que el modelo esta prediciendo un comportamiento aleatorio. Teniendo todo esto en consideración se espera que esta métrica oscile al menos entre el 0.5 y 1, considerando desde un AUC mayor a 0.8 un modelo con buen desempeño.

Finalmente, a partir de los diferentes AUC determinados para distintos modelos, es posible seleccionar a aquel que mejor desempeño presente; y de esta forma seleccionar el modelo a implementar.

Curva Precission Recall - AUCPR

La principal desventaja de la curva ROC -AUC, radica en que su enfoque esta concentrado en la clase mayoritaria, es por este motivo que también se decide observar la Curva Precission Recall.

En el eje horizontal se encuentra la sensibilidad (Recall) mientras que el vertical está la precisión, esto implica que AUC PR sea mucho más sensible a los verdaderos positivos, los falsos

positivos y los falsos negativos que el AUC [17]. Es altamente recomendado para problemas con desbalance de la data [17], ya que refleja claramente el efecto de un tratamiento de balanceo.

A diferencia del AUC este indicador no tiene un piso mínimo desde el cual es aceptable su valor, sin embargo para los diferentes tipos de clientes se define la cota mínima que esta métrica debe alcanzar para ser considerado aceptable el modelo.

Kolmogorov-Smirnov

La meétrica Kolmogorov - Smirnov, también denominada KS, representa el grado el grado de separación entre las funciones de distribución acumulativa positiva (default = 1) y negativa (default = 0) para un modelo binomial [17].

El valor del indicador KS corresponderá al punto donde se produce la máxima diferencia entre las distribuciones acumulativas mencionadas previamente, por lo que corresponde a la máxima capacidad del modelo para diferenciar la clase pagadora de la no pagadora.

Índice de Estabilidad de la población

El índice de estabilidad de la población (Population Stability Index - PSI) es una métrica que permite medir cuando ha variado la distribución de una población en el tiempo.

Se suele utilizar en el monitoreo de las características de una población con el objetivo de diagnosticar posibles problemas en el rendimiento del modelo, ya que un aumento en el psi implica cambios significativos en la distribución de la población estudiada [18].

Para calcular esta métrica se realizan los siguientes pasos [18]:

1. Ordenar la variable de puntuación en orden descendente en la muestra de puntuación (muestra inicial).
2. Dividir los datos en diferentes categorías, generalmente se suelen crear deciles.
3. Calcular el porcentaje de registros en cada grupo de la muestra de puntuación (Actual %)
4. Calcular el porcentaje de registros en cada grupo de la muestra de desarrollo (Esperado %)
5. Aplicar la fórmula para determinar el índice:

$$PSI = \sum (Actual\,\% - Esperado\,\%) \cdot ln(\frac{Actual\,\%}{Esperado\,\%}) \tag{5.11}$$

6. El criterio para interpretar el PSI es el siguiente:

- **PSI <0.1 :** sin cambios significativos en la población
- **PSI <0.2 :** cambio de población moderado
- **PSI >= 0.2 :** cambio de población significativo

Debido a la definición de la métrica se espera que Esperado % sea distinto de cero. Este error se presenta cuando la muestra de desarrollo no tiene registros en alguno de los bloques, por lo que se recomienda revisar que en todos los grupos creados se disponga de al menos un registro.

Capítulo 6

Desarrollo Metodológico

Dadas las características del trabajo a realizar se investigaron diferentes metodologías para abordar proyectos en los que se trabaja con gran cantidad de datos; algunas de las alternativas más destacadas son: SEMMA (acrónimo para Sample, Explore, Modify, Model, and Assess), Cross Industry Standard Process for Data Mining (CRISP-DM) y Catalyst. Otra metodología que suele presentarse para estos casos es Knowledge Discovery in Databases (KDD), sin embargo, las etapas que contempla este método quedan englobados en las metodologías indicadas previamente.

Se decidió utilizar CRISP-DM ya que a diferencia de las otras se plantea una constante retroalimentación entre las diferentes etapas, situación que permite avanzar o retroceder entre las diferentes etapas sin mayores inconvenientes (*véase* Figura 6.1). Además de acuerdo a Moine, J. & et.al [21], SEMMA excluye etapas como la comprensión y análisis del problema, puntos asociados al primer objetivo de este trabajo; por otro lado, Catalyst es recomendada para problemas donde el problema u oportunidad no esté definido, situación que difiere con la realidad de este proyecto.

La metodología seleccionada cuenta con seis etapas; si se considerase un avance lineal entre estas el orden respectivo correspondería a:

1. Comprensión del negocio: para lograr esto se debe evaluar el contexto del negocio y determinar los objetivos de la iniciativa.

2. Comprensión de los datos: consiste en la identificación de la información relevante para la captura inicial de datos, además de la descripción, exploración y calidad de esta data.

3. Preparación de los datos: esta fase contempla la selección, limpieza y construcción de datos; además, esta etapa debe garantizar que el formato de los datos sea acorde a la lógica del negocio.

4. Modelado: se busca seleccionar las técnicas de modelado, construir y evaluar los diferentes algoritmos seleccionados.

5. Evaluación: contempla la evaluación entregada por los modelos, además de una revisión del

proceso y la decisión sobre los futuros pasos.

6. Despliegue: para este trabajo, en esta fase únicamente se realiza la planificación del despliegue.

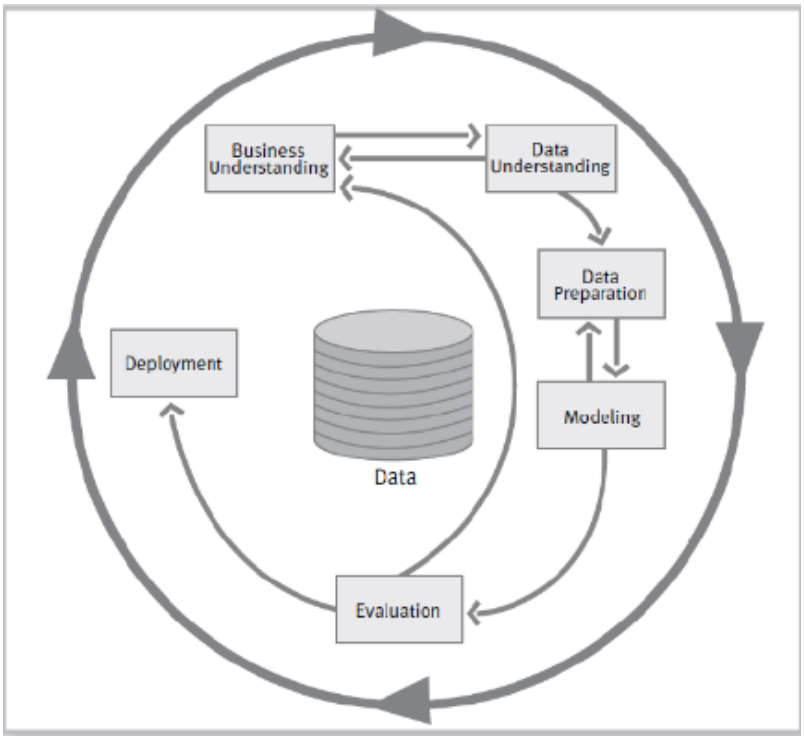


Figura 6.1: Diagrama de las diferentes etapas de metodología CRISP-DM. Extraída de documento resumen de la metodología [22]

En la documentación oficial de la metodología se plantean algunas actividades a realizar para completar cada una de las etapas, en *Anexos B* se presentan todos los objetivos de las diferentes etapas, además de actividades para lograr esas metas.

6.1. Comprensión del negocio

Se estima que los ingresos orgánicos de la industria de telecomunicaciones en Chile durante 2021 crecieron un 7,4 % con respecto al año 2020 [2]. Este aumento se debe esencialmente a dos líneas de negocio: la primera corresponde al sector móvil, -mediante la venta de equipos y servicios móviles-; mientras que la segunda línea corresponde a servicios fijos, destacando a los productos de internet fija.

En línea con estos antecedentes la compañía ha decidido invertir cerca de US$660 millones durante el 2022 con un sólido foco en los productos móvil y hogar. Es la mayor inversión desde 2018, cuando el monto ascendió a US$680 millones [19]. De esta gran inversión, se tiene completado destinar cerca de US $106 millones al negocio hogar, el objetivo de estas inversiones radica en el despliegue de fibra óptica y la conexión con nuevos clientes del segmento [20].

Es importante destacar, que este plan estratégico ha impulsado a la línea hogar como una prioridad de la empresa; sin embargo, es relevante indicar que esta decisión surge a raíz del nuevo escenario cultural de Chile producido por la pandemia del COVID 19. Esta última ha conllevado a nuevas formas de realizar los trabajos, antes del 2020 eran pocos los trabajadores que contaban con modalidades híbridas u online; situación contraria al escenario del 2022, donde ha crecido el porcentaje de personas que trabajan o estudian desde sus hogares. Es evidente la presencia de un cambio cultural, ya que este comportamiento ha permanecido vigente incluso sin la obligatoriedad de aislamiento; por supuesto, esto ha implicado un aumento de la exigencia de los consumidores de estos productos, ya que se ha vuelto más relevante el acceso a una conexión más rápida y sin interrupciones.

El creciente aumento de la demanda por los servicios de internet y la expansión de fibra óptica de la compañía, conlleva a la necesidad de evaluar a los potenciales clientes que buscan acceder a algún producto de la línea. Cuando un cliente quiere contratar un producto puede acudir a los diferentes canales dispuestos por la empresa para consultar la factibilidad de su requerimiento; esta solicitud permite a la empresa realizar consultas financieras a bureaus; además, de investigar si la dirección del cliente cuenta con cobertura para el servicio que solicita.

Una vez la solicitud es aprobada e instalado el servicio, el cliente recibe a finales de ese mes la primera factura la cual generalmente corresponde a un monto proporcional de los días que tuvo activo el servicio. En la figura 6.2 se ejemplifica el caso de un cliente que pertenece a la camada de Agosto del 2020, esta persona a finales de Septiembre recibirá su segunda factura la cual esta vez contiene el total del plan contratado.

Ya que la segunda factura corresponde al monto real del plan, el foco se encuentra en esta boleta. Una vez determinada la factura con la cual se trabaja es relevante definir la ventana de tiempo en que se observa el comportamiento de pago: se dispone de vistas a 30 y 60 días, la primera no es un buen indicador de la propensión al no pago ya que un leve retraso puede enmarcar a clientes como malos pagadores; sin embargo, si 60 días después de la segunda factura un cliente continúa sin regu-

larizar su situación esto revela que efectivamente no existe una intencionalidad de pagar el servicio.

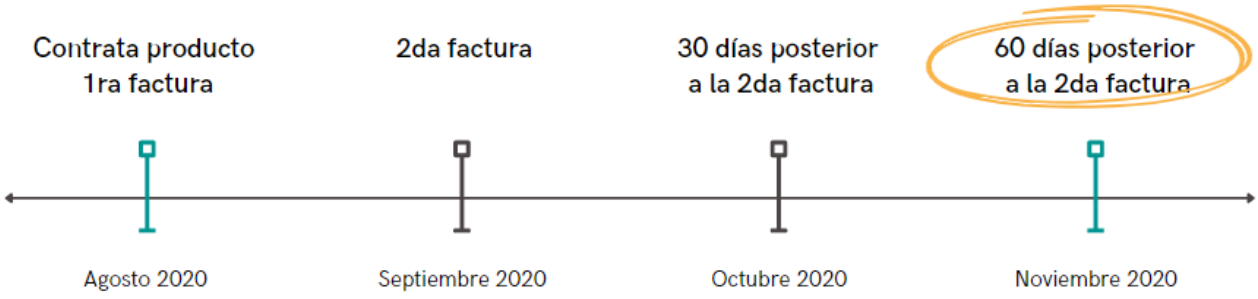


Figura 6.2: Ejemplificación facturación de clientes y construcción de la variable objetivo. Elaboración propia.

Por lo que la QNP y TNP objetivo corresponden a las **tasas 60 días posterior a la segunda factura**. Para el ejemplo graficado en la Figura 6.2 la camada corresponde a Agosto 2020, las tasas QNP y TNP podrán ser determinadas en Noviembre 2020 con los clientes que a esa fecha posean un saldo mayor a cero.

Una vez determinada la variable que representa el efecto del problema, se procede a graficar las tasas que evidencian el comportamiento estudiado para los diferentes perfiles de clientes.

QNP

En las Figuras 6.3, 6.4 y 6.5 se evidencia un aumento de la cantidad de clientes facturados en fibra en el 2021 con respecto al año 2020, independiente de la antigüedad del cliente con la empresa.

En cuanto a la QNP, para los clientes nuevos y semi nuevos se observa un aumento considerable durante Agosto del 2020 hasta Diciembre del 2020. Esta situación comienza a regularizarse durante Enero del 2021, pero pese a esto es clara la tendencia al alza de la QNP. Respecto a los clientes antiguos, también se observa una tendencia al alza de la QNP, pasando de una tasa 0,63 % en Agosto del 2020 a una tasa de 2,37 % en Diciembre del 2021.

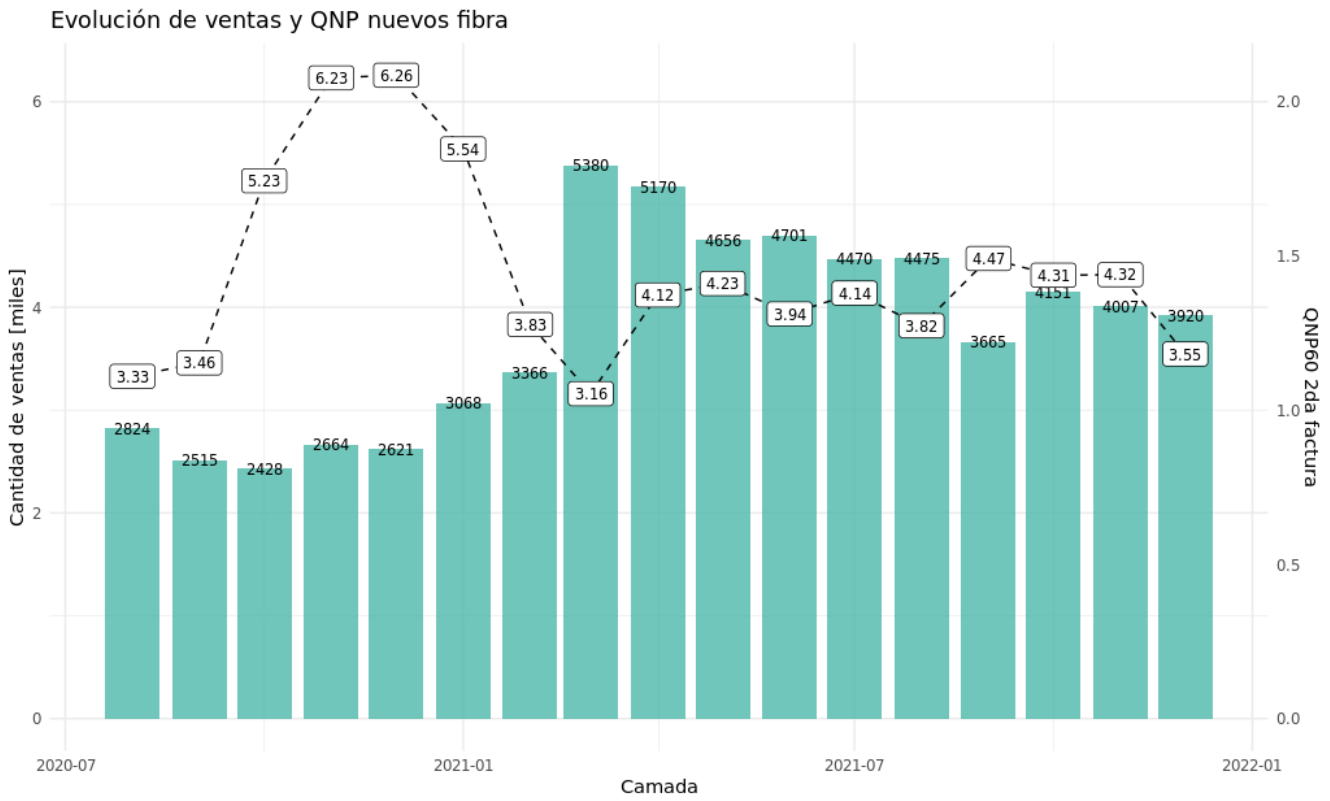


Figura 6.3: Evolución tasa QNP 60 días posterior a la segunda factura para clientes nuevos fibra.

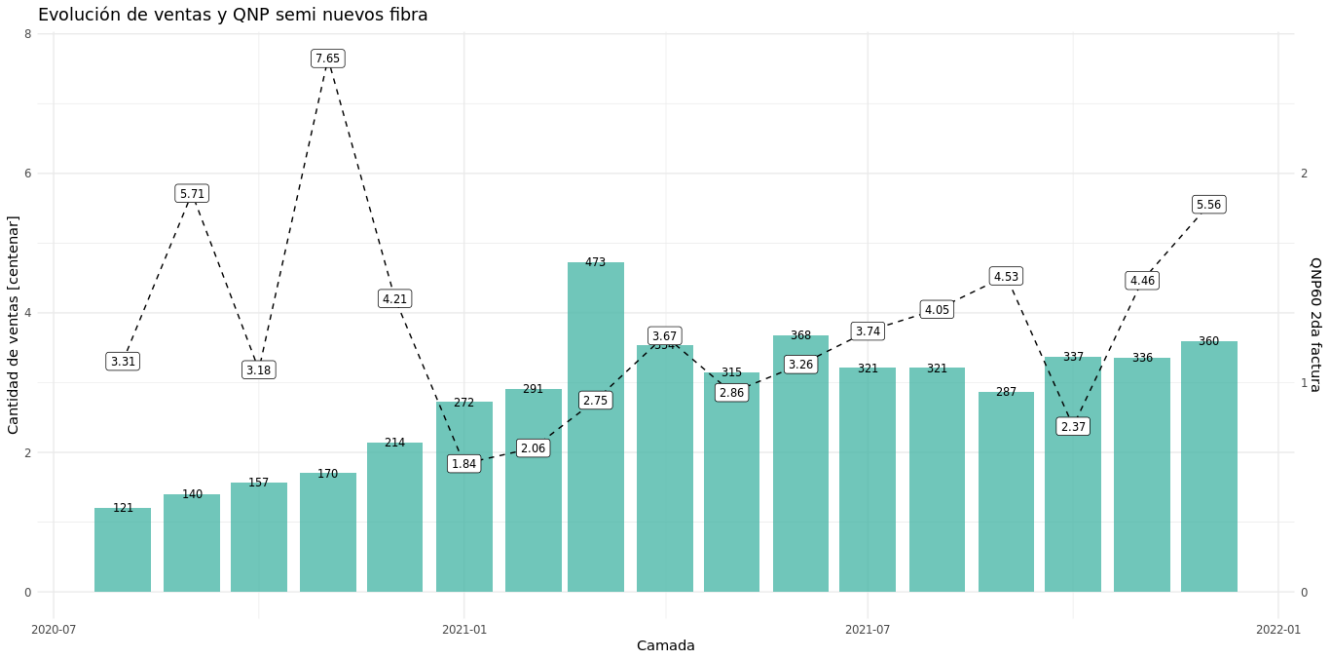


Figura 6.4: Evolución tasa QNP 60 días posterior a la segunda factura para clientes semi nuevos fibra.

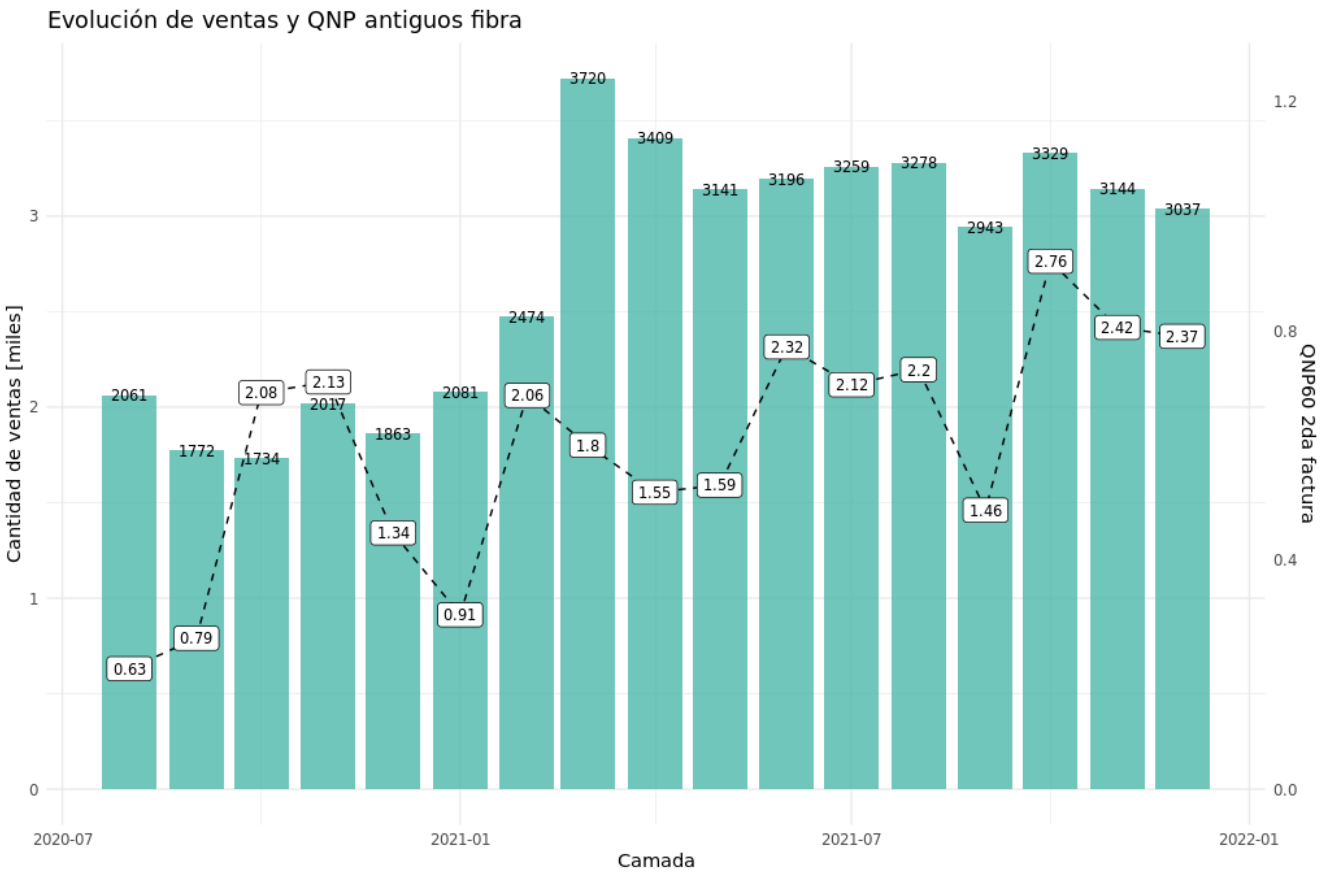


Figura 6.5: Evolución tasa QNP 60 días posterior a la segunda factura para clientes antiguos fibra.

TNP

Del apartado previo se evidencia un aumento en la cantidad de ventas, transversal a los perfiles de clientes. En las Figuras 6.6, 6.7 y 6.8 se observa la evolución de la TNP a 60 y 120 días posterior a la segunda factura para los diferentes segmentos de clientes definidos previamente. De estos gráficos es posible inferir que al menos desde Enero del 2021 se observa una tendencia al aumento en las TNP para el servicio fibra independiente de la antigüedad del cliente.

Adicionalmente, en todos los segmentos de clientes, se observa un comportamiento anómalo en las tasas de no pago durante el periodo Agosto 2020 a Enero 2021. Posterior a esa ventana, las TNP comienzan a tener una conducta más estable pero siempre con una clara tendencia al alza.

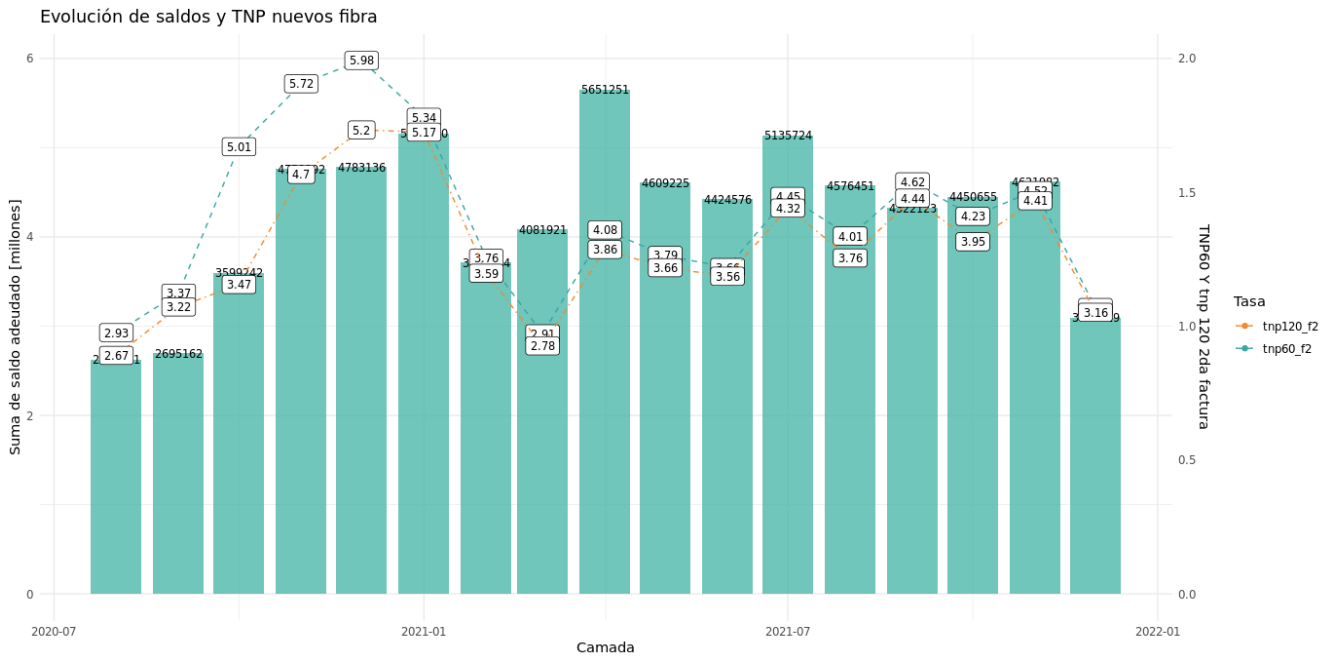


Figura 6.6: Evolución TNP 60 días posterior a la segunda factura y TNP 120 días posterior a la segunda factura, de clientes nuevos fibra.

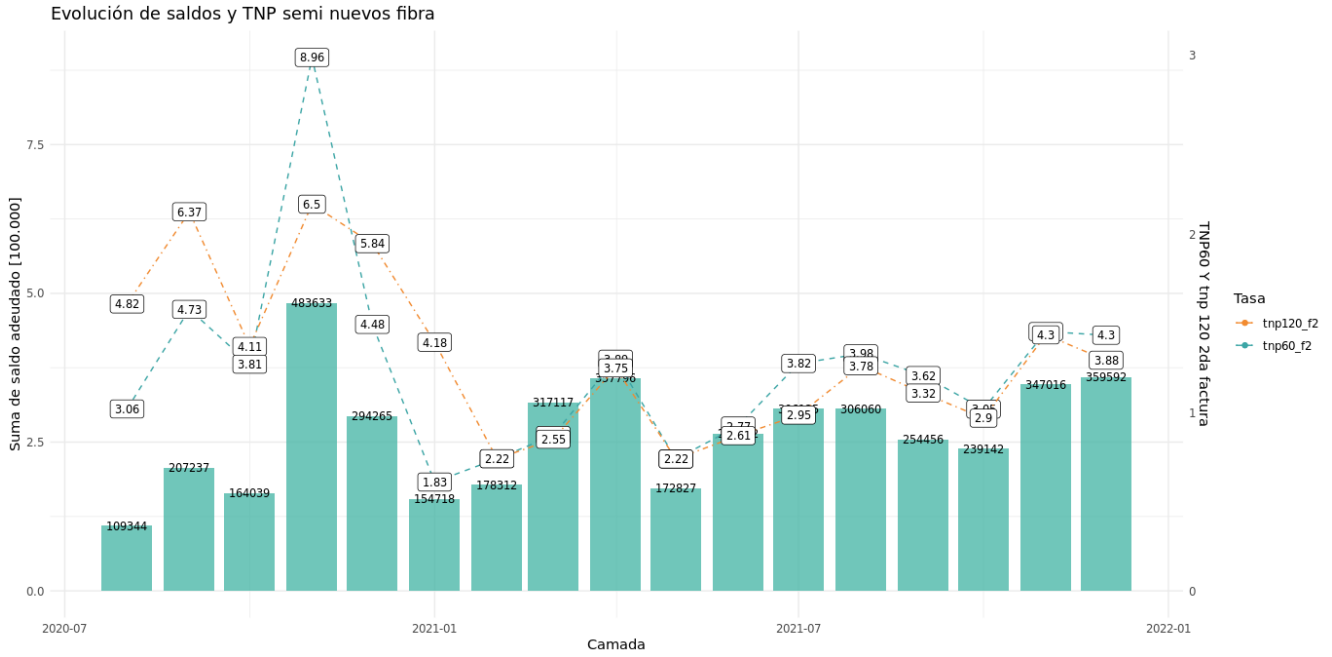


Figura 6.7: Evolución TNP 60 días posterior a la segunda factura y TNP 120 días posterior a la segunda factura, de clientes semi nuevos fibra.

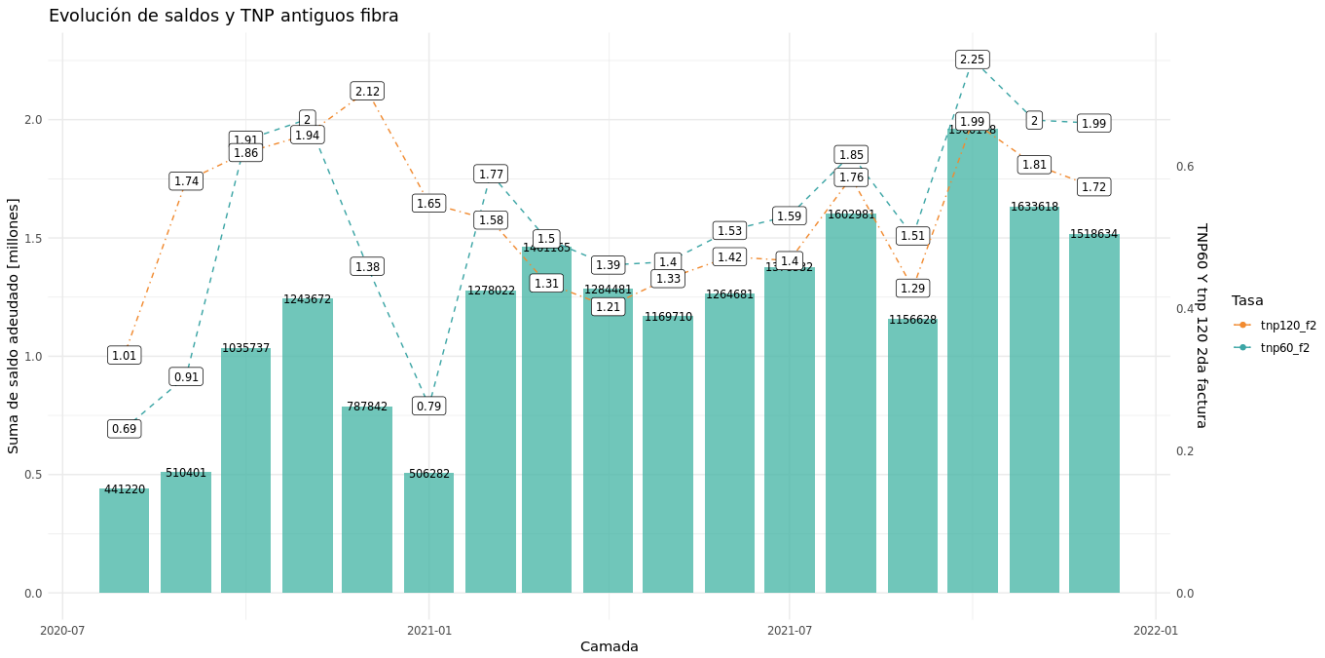


Figura 6.8: Evolución TNP 60 días posterior a la segunda factura y TNP 120 días posterior a la segunda factura, de clientes antiguos fibra.

6.2. Comprensión de los datos

La captura inicial de los datos proviene de dos fuentes de información: la tabla que registra las **ventas de hogar** y la que registra todas las **consultas realizadas a los bureaus financieros** para determinar la entrega de un servicio, ambas están almacenadas en el datastore de la empresa y su actualización está a cargo del equipo de Big Data. Para acceder a los registros basta con realizar querys al datastore con la información requerida.

Para esta sección se recomienda la revisión de las definiciones claves dispuestas en el segundo capítulo.

6.2.1. Ventas hogar

Esta tabla tiene por objetivo almacenar información relacionada a las ventas de los diferentes servicios de la línea hogar de la compañía; posee registros que comprenden desde marzo del 2020 a agosto del 2022, disponibles en 46 variables. Esta tabla es de especial interés ya que a partir de las columna saldo60_f2 es posible construir la variable **default**, mediante la cual es posible determinar las QNP y TNP.

Las 46 variables que dispone esta base de datos se pueden caracterizar en los siguientes grupos, donde algunas de las variables más representativas se presentan a continuación:

- **Información de venta:** se encuentra variables como la fecha, año, camada e identificador asociado a la venta.
- **Contexto venta:** canal, sucursal e información del agente que concreta la venta.
- **Características del cliente:** rut anonimizado, región donde habita e indicadores de portabilidad.
- **Servicio contratado:** incluye variables como el tipo de servicio contratado, el tipo de plan, monto del plan, primera y segunda factura, cantidad de decos, etc.
- **Comportamiento cliente:** saldos pendientes del cliente en diferentes ventanas de tiempo (30 y 60 días) con respecto a diferentes facturas (primera y segunda factura).

  Estas columnas se van completando con el transcurso del tiempo dependiendo del comportamiento de pago del cliente.

  En esta categoría también se incluyen tres indicadores del nivel de riesgo del cliente.
- **Antigüedades:** Variables categóricas que indican la antigüedad del cliente en diferentes mercados y productos, algunas de estas son: antigüedad móvil, antigüedad hogar, antigüedad en el sistema interno, entre otras.

En términos generales, de las 46 variables que se dispone: 4 son variables de fecha, 8 variables de caracteres (7 de ellas en realidad corresponden a variables categóricas), 21 son variables categóricas y toda la diferencia (13) corresponden a variables numéricas. Adicionalmente, **todas las columnas cuentan con más del 95 % de completitud de los datos.**

Hallazgos relevantes

El principal insight obtenido de esta tabla corresponde al comportamiento de los diferentes servicios dada la antigüedad del cliente (*véase* Figura 6.10 y 6.9). A partir de los gráficos de las tasas de no pago, se logró evidenciar que TV light cuenta con el comportamiento más riesgoso de la línea, ya que sus tasas de no pago presentan amplias oscilaciones y no logran evidenciar una tendencia al alza o a la baja.

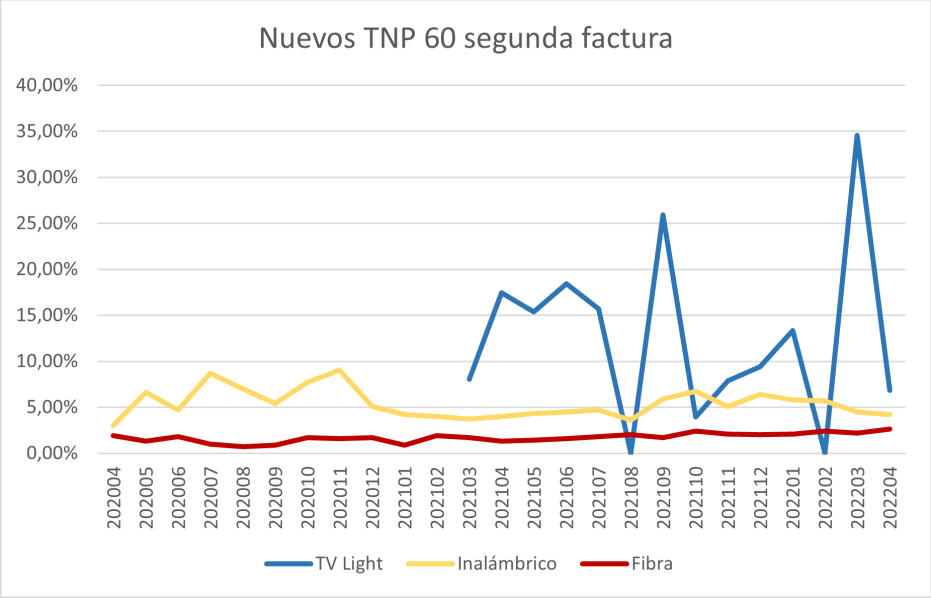


Figura 6.9: Evolución de las TNP 60 días posterior a la segunda factura, para clientes nuevos durante el periodo comprendido entre Abril 2021 a Abril 2022.

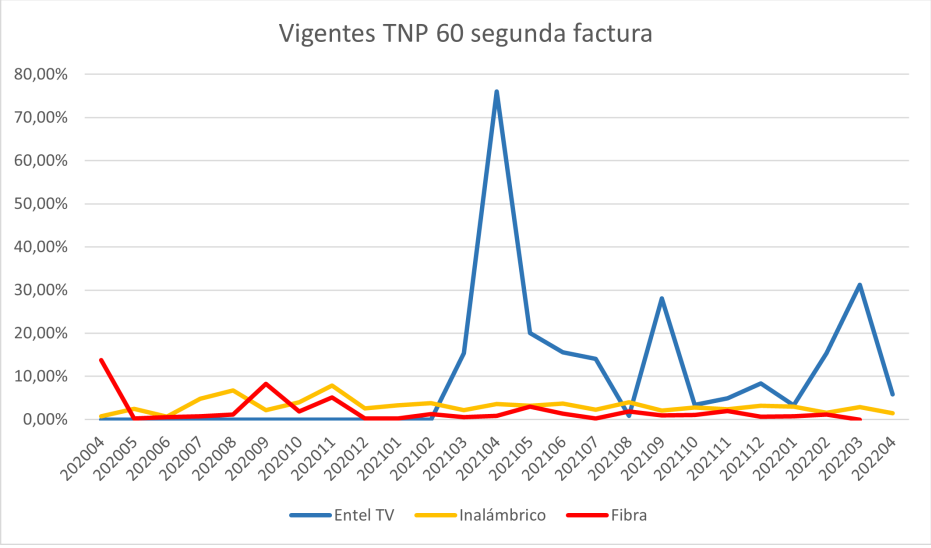


Figura 6.10: Evolución de las TNP 60 días posterior a la segunda factura, para clientes semi nuevos y antiguos durante el periodo comprendido entre Abril 2021 a Abril 2022.

Por el contrario, el servicio con mejor desempeño corresponde a fibra, -para toda clase de clientes-, siendo el segmento de clientes nuevos aquel que posee una TNP levemente superior pero aún con posibilidad de mejora.

En lo que respecta a la distribución mensual de ventas desagregada por producto es posible indicar que existe una baja en el servicio inalámbrico durante el año 2021 (*véase* Figura 6.11); fibra ha tomado parte de ese segmento de ventas que inalámbrico pierde, mientras que TV se ha mantenido estable durante la misma ventana temporal .

Ahora bien, al profundizar en los diferentes paquetes disponibles (mono, dúos o tríos) se observa que únicamente cerca del 20 % corresponden a packs, -ya sean dúos o tríos-, (*véase* Figura 6.12), por lo que el grueso de la venta de productos hogar radica en la venta de mono paquetes, siendo los productos de internet los más cotizados.

Ahora bien, del total de la muestra estudiada se logró determinar que el 43 % de estos clientes corresponden a **clientes completamente nuevos**; es decir, del total de los clientes nuevos el 43 % corresponde a personas que no poseían un contrato móvil activo con la empresa. La diferencia restante (57 %) corresponderá a clientes nuevos en servicios hogar pero que si disponen de un contrato activo en la categoría móvil.

Finalmente, se observa que los productos más contratados por los clientes nuevos corresponden a los servicios inalámbricos y fibra (*véase* Figura 6.13).

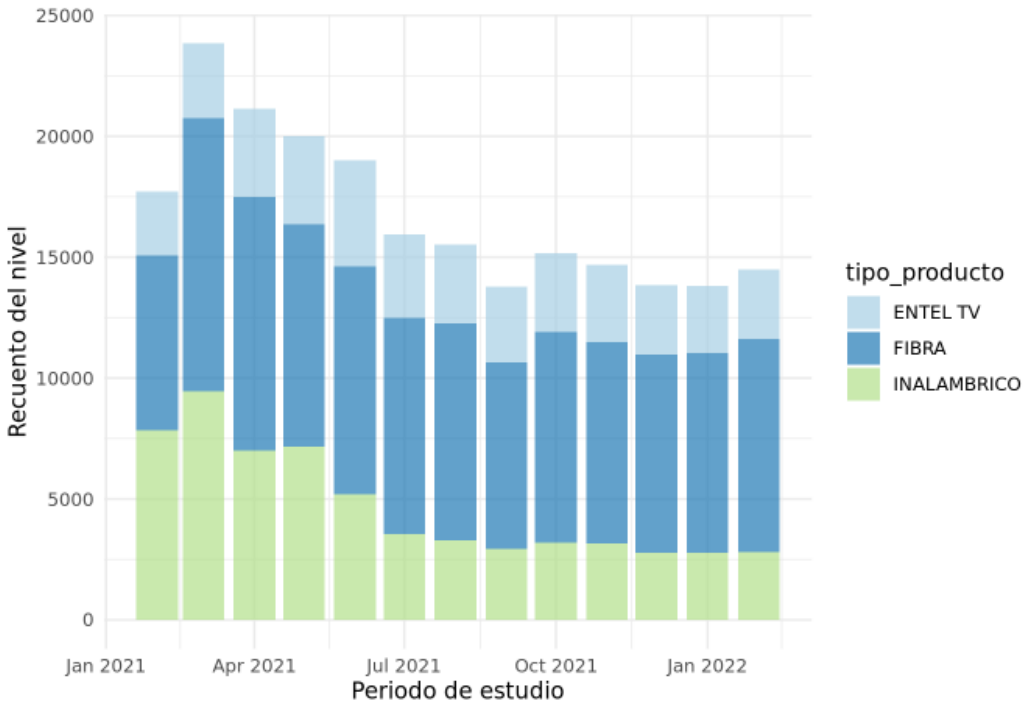


Figura 6.11: Distribución mensual de productos, desagregado por servicio de abertura.

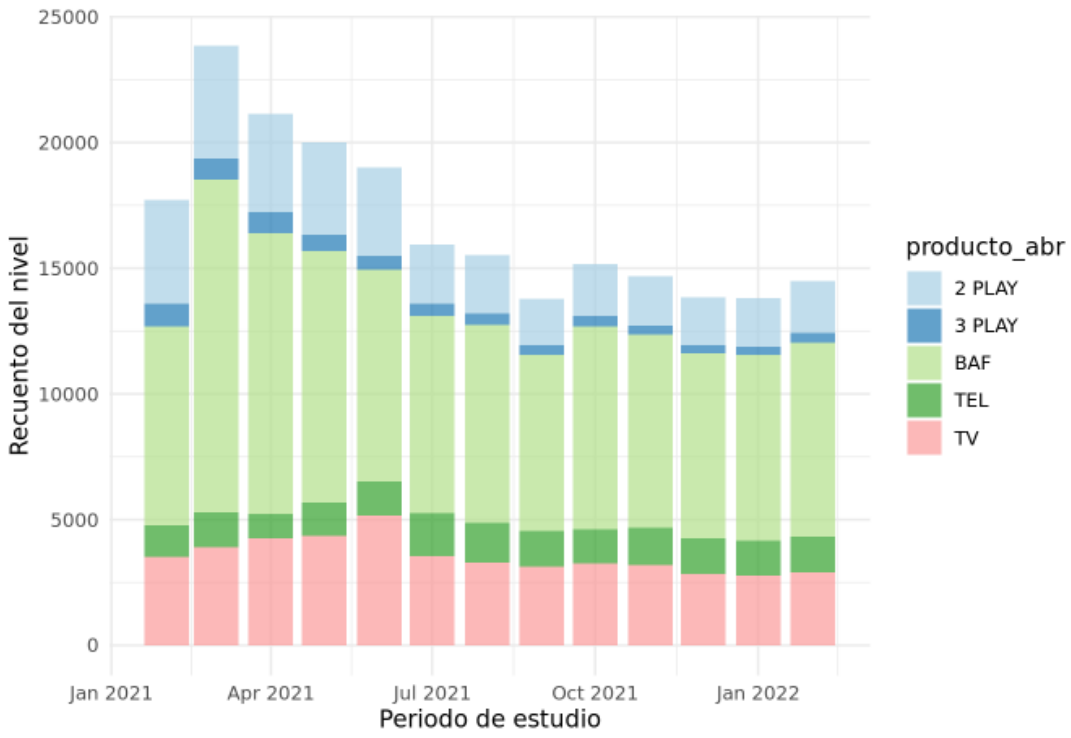


Figura 6.12: Distribución mensual de productos, desagregado por tipo de plan/paquetización vendida.

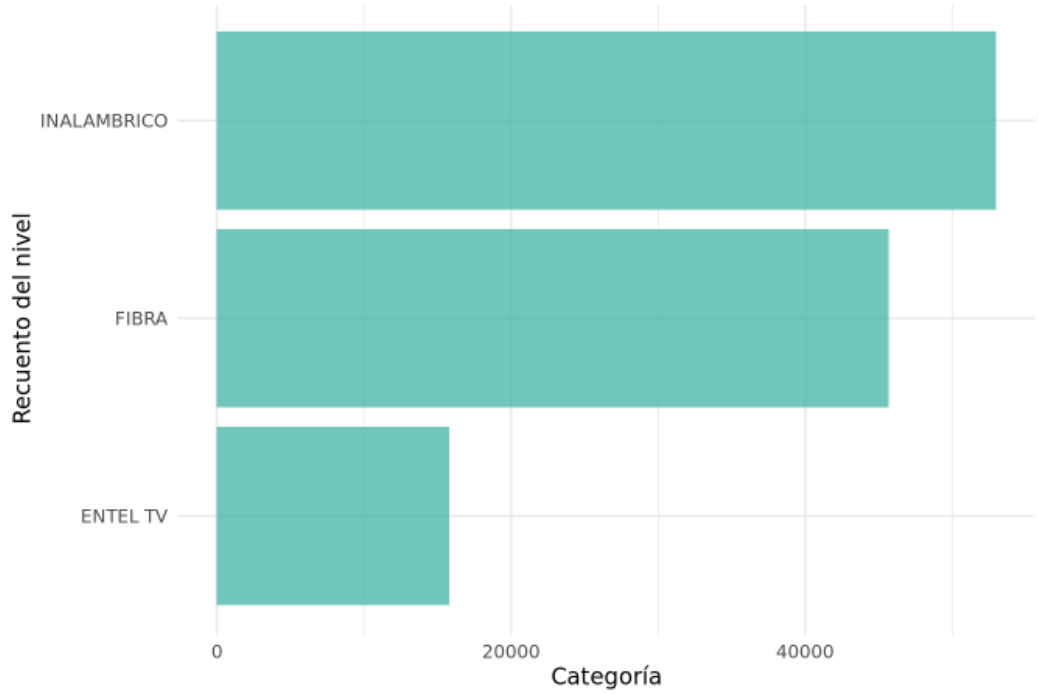


Figura 6.13: Preferencias de productos para los clientes nuevos.

6.2.2. Consultas financieras (PCO)

Cuando a una persona quiere contratar un servicio, esta acude a un canal de contacto con la empresa; al levantar su solicitud la compañía solicita diferentes datos de la persona, y partir de estos últimos realiza consultas a diferentes fuentes de información con el objetivo de clasificar el riesgo de la personas y/o las diferentes alternativas de productos que podrían ser contratados por la persona acorde a su requerimiento.

Esta tabla almacena todas las evaluaciones financieras realizadas, además de si estas fueron aprobadas o rechazadas; los orígenes de información para la construcción de la evaluación son esencialmente tres:

- **Bases externas:** data recopilada y entregada por bureaus financieros. De este origen, proviene esencialmente el comportamiento financiero del potencial cliente con otras empresas; algunas de las variables con las que cuenta son: monto morosidades e impagos con diferentes ventanas de tiempo, bancos en los que registra movimiento, avalúos fiscales, entre otras.

  También entrega información más particular del solicitante como: estado cédula de identidad, actividad económica, profesión, edad, estado civil, cantidad de direcciones registradas, etc.

- **Bases internas:** información disponible de la persona en los registros internos de la compañía, estos campos solo cuentan con información si es que el cliente posee productos de otro mercado o si en el pasado obtuvo acceso a algún producto.

  Además de las respectivas antigüedades y planes a los que accedió, se encuentran variables del tipo: cantidad de facturas, facturas canceladas, montos adeudados con la empresa, líneas actuales, días de mora, etc.

- **Otros:** corresponden esencialmente a resultados de la evaluación, además de información recopilada en la solicitud del cliente, tales como: fecha, identificador y canal de solicitud de evaluación, estado de la solicitud, fecha de actualización, y demás.

En términos globales, cuenta con 1296 variables y con registros que abarcan evaluaciones realizadas desde Marzo del 2020 a Septiembre del 2022. Para el análisis exploratorio se utiliza una muestra compuesta por 1.735.066 registros, todos asociados al periodo de diciembre del 2021; es importante destacar que en esta fuente se almacenan todos las consultas realizadas por lo que se requiere filtrar por tipo de negocio (en este caso hogar).

Una vez seleccionadas las evaluaciones de interés, se procede a unificar esta información con la tabla de ventas hogar con el objeto de rescatar aquellas consultas que efectivamente se tradujeron en una venta, y así conocer el historial de pago asociado al cliente de una determinada consulta. Debido a que no se cuenta con un identificador que indique la consulta que se transformo en venta, al momento de hacer el cruce se considera la última consulta aprobada para un determinado cliente durante la misma camada en la que se concreto una venta.

Hallazgos relevantes

Mediante los primeros dos puntos expuestos en el marco conceptual para la selección de variables, se comienza a explorar con un tablón el cual estaba compuesto de la siguiente forma: 40 % corresponde a fuentes externas, 37 % a fuentes internas y la diferencia corresponde a la categoría otros definida previamente.

Del total de variables, 9 corresponden a variables categóricas (almacenadas como caracteres), como: estado civil, profesión, estado cédula de identidad, clasificación de clientes, entre otras. El resto corresponde únicamente a variables numéricas (almacenadas como integer64). En lo que respecta a los missing values, 608 variables (equivalente al 47 % de las variables totales) contienen más de un 80 % de NA; estas variables fueron removidas para la comprensión de la data ya que no serían utilizadas de todos modos.

Acorde a la metodología de selección de variables propuesta en el marco conceptual, se procede a calcular el Information Value de las 688 variables restantes con respecto a la variable objetivo.

La mayoría de las columnas corresponden a la cantidad de días, montos o cantidad de documentos de variables como protestos, morosidad e impagos a determinadas ventanas de tiempo. Esto va en concordancia con la lógica supuesta, ya que se espera que dado que una persona posee una mayor cantidad de días morosos esta persona tenga una mayor propensión al no pago de servicio. También se cuenta, aunque minoritariamente, con variables que registran el efecto contrario al recientemente expuesto, un caso corresponde a la variable que indica la cantidad de bienes raíces del consultante, ya que a mayor cantidad bienes raíces se observa una menor tasa de no pago, debido obviamente a que se esa persona dispone de los medios económicos para pagar el servicio sin mayores inconvenientes.

6.3. Preparación de los datos

6.3.1. Construcción del tablón

A continuación se presentan todos los pasos realizados para la construcción del tablón final con el que se realizan los modelos.

1. Construir join query que extraiga de la tabla ventas hogar: rut anonimizado del cliente, la camada de ventas, tipo de servicio vendido y el saldo 60 días posterior a la segunda factura. De la tabla que contiene las consultas financieras se solicitan las 49 variables relevantes levantadas en la sección anterior, además de algunas variables de identificación de la consulta como: rut anonimizado de la persona que solicita la evaluación, el resultado de esta (si fue aprobada o rechazada) y el mes en el que se realiza.

   La unión de la consulta se realiza imponiendo la igualdad de los ruts anonimizados de los clientes y la camada de ventas con el mes de realización de la consulta.

   Finalmente, a esta query se le impone la restricción del tipo de negocio, es decir que solo tome consultas realizadas para servicios del mercado hogar.

2. Se tabulan la cantidad de ventas obtenidas agrupadas por camadas (*véase* Figura 6.14), a partir de la cantidad de datos se determina que la ventana de tiempo que contendrá el tablón corresponde desde Agosto 2020 a Mayo del 2022 debido a que durante estos meses se cuenta con una cantidad estable de registros; además de que la última tasa de no pago 60 días segunda factura disponible corresponde a la tasa de no pago de la camada de Mayo de este año (la cual recién fue determinada a finales de Agosto).

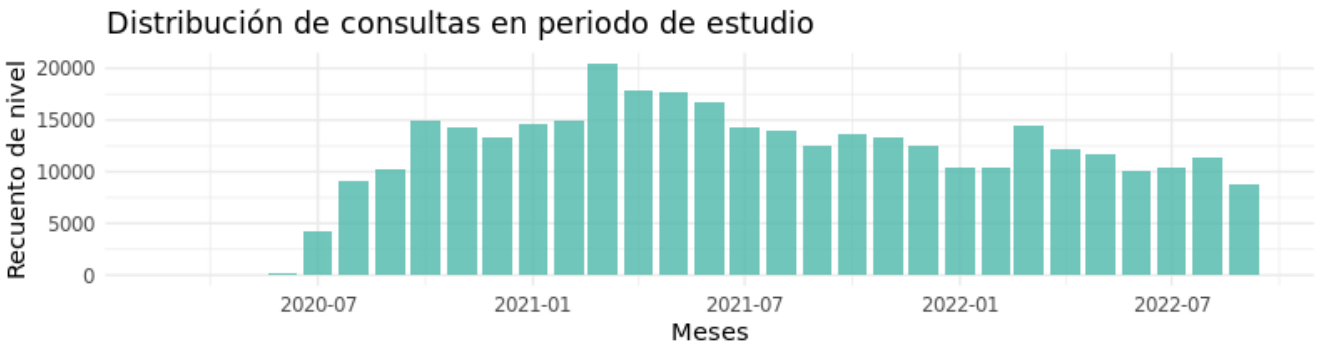


Figura 6.14: Distribución de personas que realizaron consultas diferenciadas por camadas

3. El tercer paso consiste en la construcción de la variable objetivo 'default' a partir de la variable saldo 60 segunda factura (esta última es removida del tablón).

4. Reconocer todas las variables integer64 como numéricas.

5. Mapear la cantidad de missing values, se obtuvo que ninguna de las variables contenían más de un 80 % de NA's, por lo que ninguna fue removida.

6. Se construyen dos variables para rescatar el historial de consultas: la primera fue denominada 'qconsultas' y consiste en crear un contador todas las consultas registradas para un cliente

independiente de si fueron aprobadas o no. En paralelo se contruye 'qconsultas aprobadas' que tiene por objetivo almacenar la cantidad máxima de consultas aprobadas para el cliente.

7. Se seleccionan los registros que tengan la consulta más reciente acorde a la camada de venta del servicio y remueven los registros duplicados del set de datos.

8. Se itera nuevamente information value para ver si alguna variable es removida, el resultado obtenido indica que las variables 'qconsultas' y 'qconsultas aprobadas' no son buenos predictores de la variable objetivo.

9. Se procede a graficar el efecto de todas las variables provenientes de las consultas financieras en la variable objetivo, para esto se utiliza la transformación WoE Binning (*véase* Marco conceptual sección Selección de variables). Este método transforma las celdas en blanco en NA's, y además agrupa a todos los missing values en una categoría; los gráficos y sus respectivos análisis se presentan en la siguiente sección.

10. Eliminar del tablón las variables que no son capaces de agrupar comportamientos de no pago, desglose en la siguiente sección.

11. Finalmente, el tablón resultante del procesamiento consta con: **44 variables** de las cuales 4 están relacionadas a la venta y 40 a la información de consultas. Los **registros** para la ventana temporal definida previamente alcanzan **303.367**.

6.3.2. Análisis de las variables mediante WoE

Para profundizar en el impacto del comportamiento de las variables asociadas a las consultas financieras dado default se utilizó WOE Binning.

De este análisis se logra indicar que las siguientes variables deben removidas ya no son capaces de detectar comportamientos relevantes para la variable default:

- bases externas bureau monto protestos ultimos 24 meses
- bases internas comportamiento cliente sparelibre num2
- bases externas bureau monto protestos mayor 36 meses
- bases externas bureau monto protestos ultimos 12 meses
- bases internas comportamiento cliente monto mora interna
- bases externas bureau monto protestos ultimos 6 meses
- bases externas bureau monto protestos ultimos 3 meses
- bases internas comportamiento cliente sparelibre num1

También se removerá bases externas bureau score, debido a que se desconoce como se construye este score.

Todas las variables asociadas a impagos y montos poseen un comportamiento bastante similar

Dado que se cuenta con muchas variables se decide profundizar en el comportamiento de las variables más relevantes. En lo que respecta a la **morosidad**, se observa en las Figuras 6.15 y 6.16 que las curvas de QNP de cantidad y monto de morosidades poseen comportamientos similares. Los valores faltantes poseen una QNP cercana al 15 % y corresponden a la categoría con la QNP más alta de personas que no pagan el servicio.

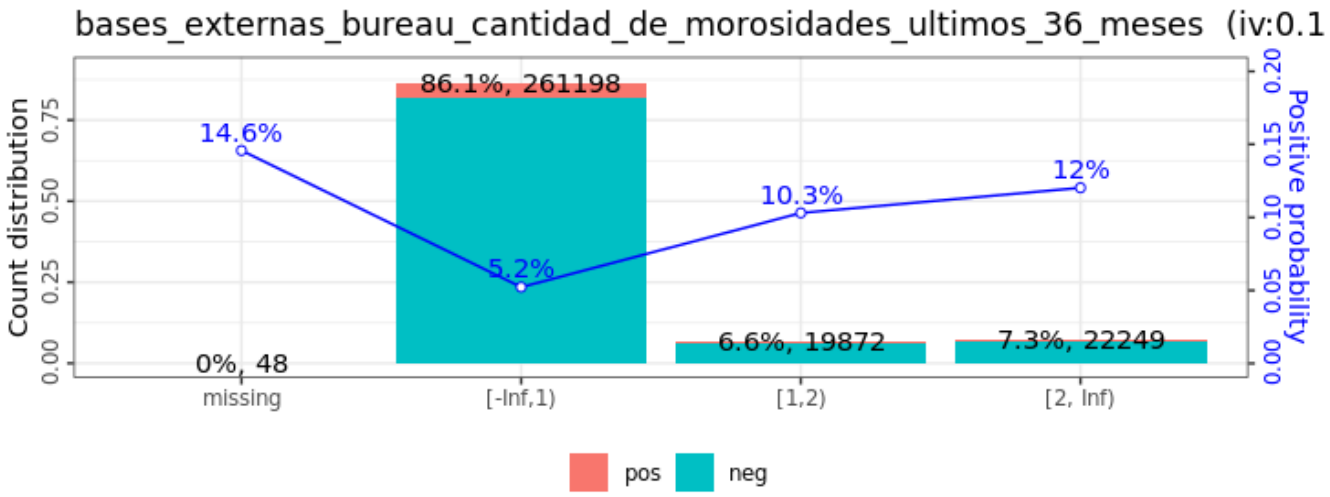


Figura 6.15: QNP agrupada para variable cantidad de morosidades últimos 36 meses.

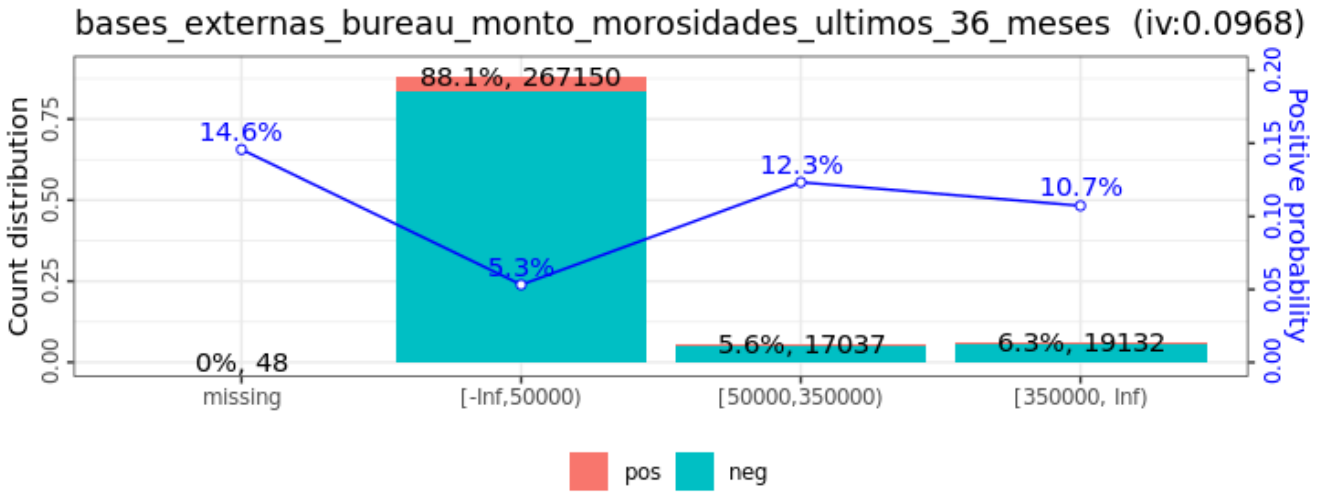


Figura 6.16: QNP agrupada para variable monto morosidades últimos 36 meses.

Dos variables asociadas a los **impagos** se presentan en las Figuras 6.17 y 6.18. De estos gráficos es posible observar que la cantidad y monto de impagos posee un comportamiento bastante similar a morosidades, por lo que es una alarma para el estudio de las correlaciones entre estas variables. Los valores faltantes nuevamente representan una QNP cercana al 15 %, y responden a la categoría con la QNP tasa más alta.

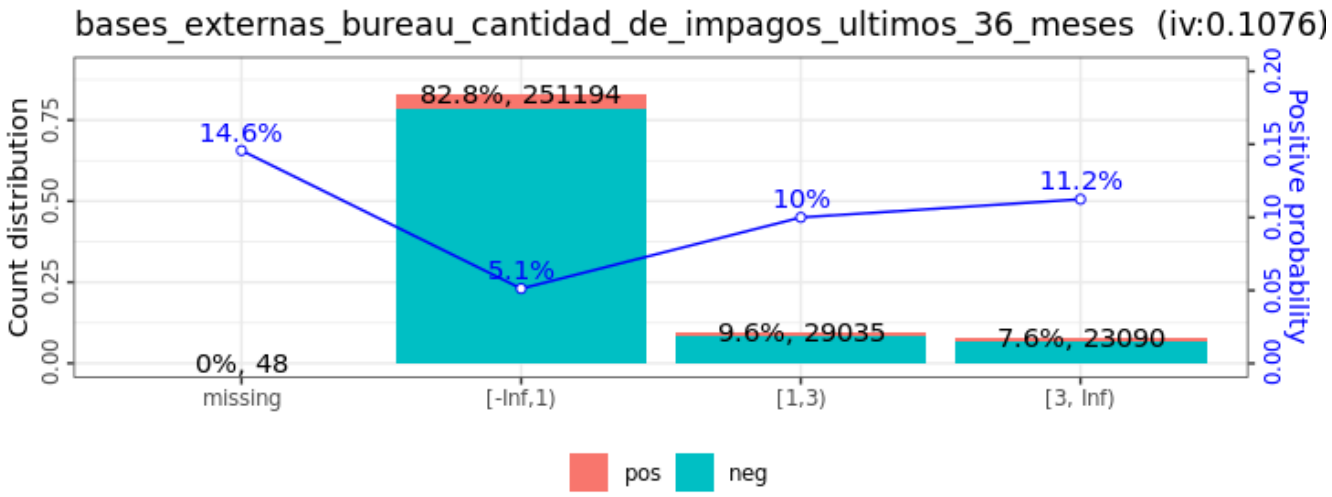


Figura 6.17: QNP agrupada para variable cantidad impagos últimos 36 meses

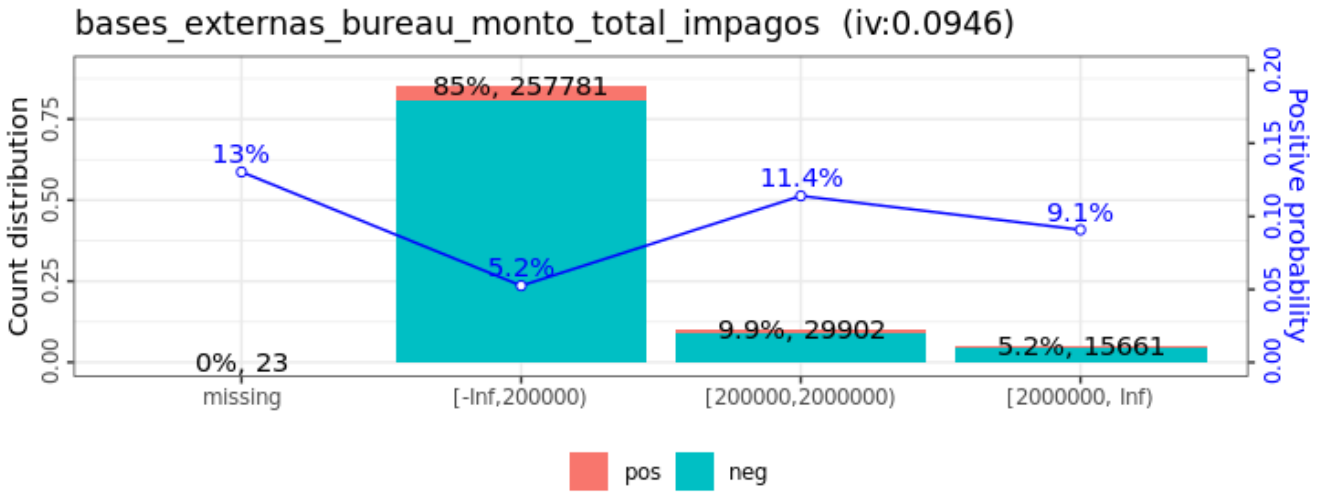


Figura 6.18: QNP agrupada para variable monto total impagos

Una variable interesante de observar es la cantidad de bancos en que un potencial cliente ha registrado movimientos, ya que esta variable es incorporada en todos los modelos finales (*véase* Figura 6.19). Nuevamente, la categoría que presenta la mayor cantidad de clientes que no pagan el servicio corresponde a los "missing values". Luego las personas que solo han registrado movimientos en 0 o 1 banco, corresponden a la segunda mayoría, obteniendo una QNP cercana al 7 %.

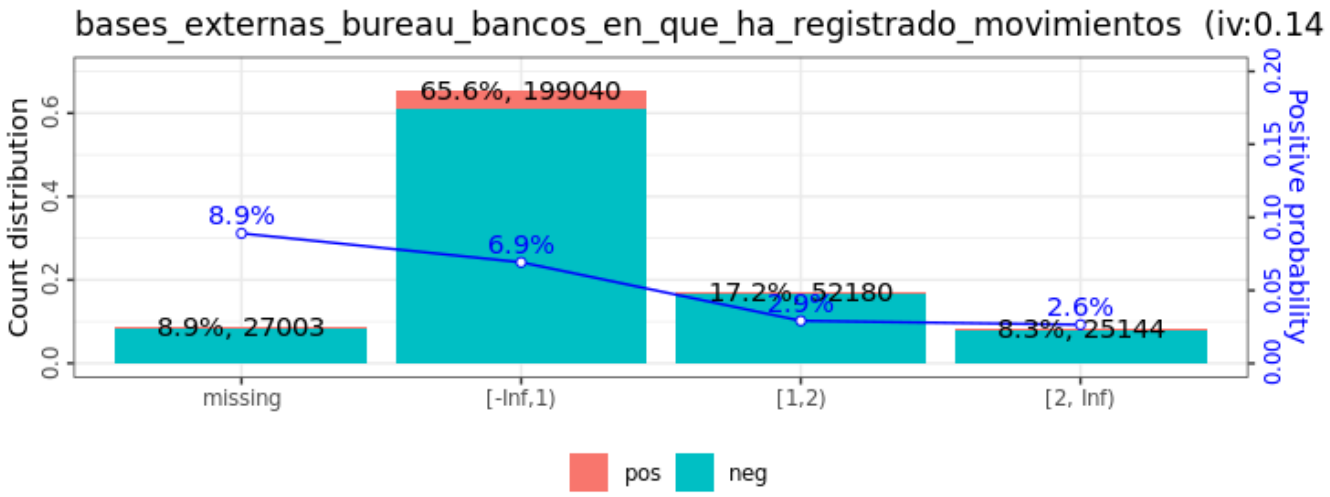


Figura 6.19: QNP agrupada para variable bancos en que ha registrado movimientos

EL comportamiento de las categorías del índice socio económico de emergencia (ISE) se expone en la figura 6.20. Se observa que los clientes potenciales que pertenecen a los grupos E y D poseen las QNP más altas. Las QNP aumentan, a medida que el potencial cliente es clasificado en un grupo más vulnerable.

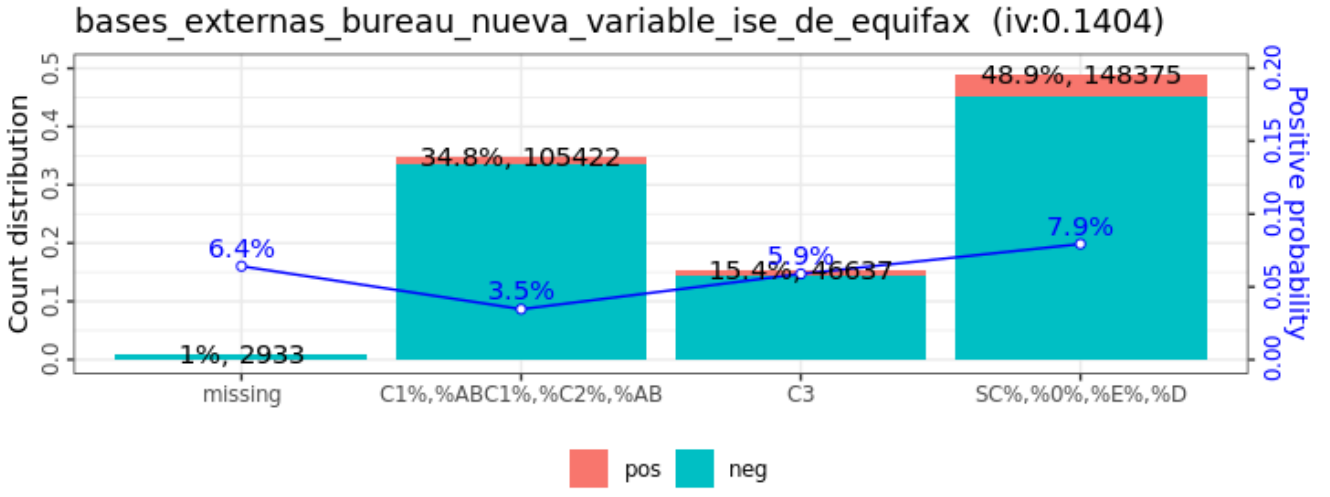


Figura 6.20: QNP agrupada para variable Índice socio económico de emergencia.

El comportamiento de la variable cantidad de facturas (*véase* Figura 6.21), indica que el riesgo suele estar concretado durante las primeras 15 facturas (cerca de un año). Posterior esa cantidad la QNP comienza a decaer considerablemente, lo cual tiene sentido ya que a medida que un cliente pague más facturas es muy probable que sea un buen pagador.

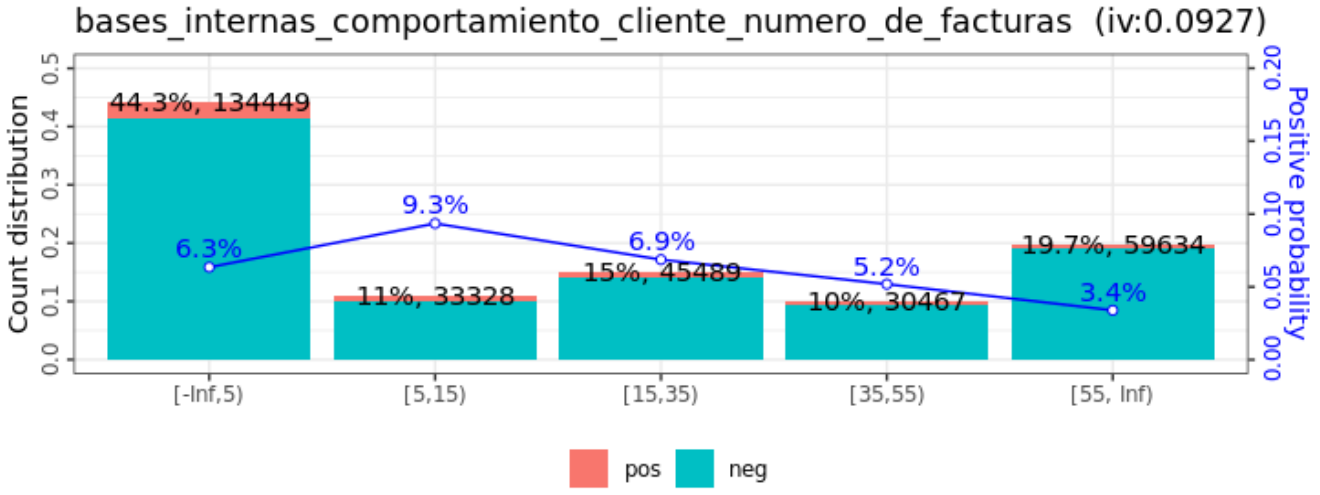


Figura 6.21: QNP agrupada para variable interna número de facturas.

Los días de mora interna se grafican en 6.22, esta variable es muy importante porque evidencia que a medida que aumentan los días de mora lo hace la QNP. Es posible inferir de la figura que si un cliente presenta 9 días o más de mora es más probable que no pague el servicio que uno que presenta 2 días de mora.

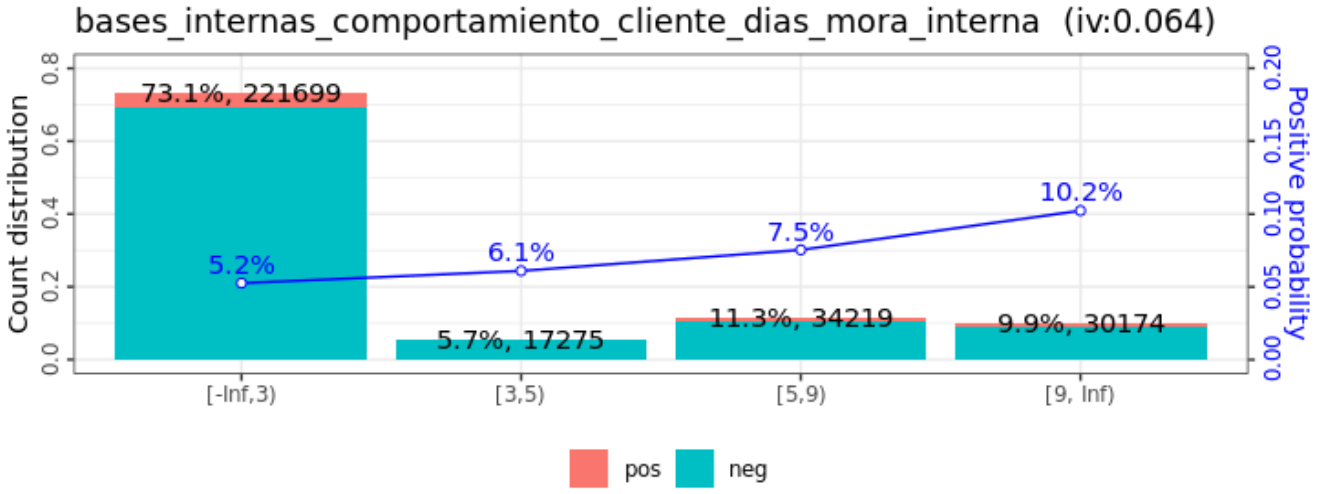


Figura 6.22: QNP agrupada para variable días de mora interna.

La última variable en la que se profundiza corresponde al monto promedio que los clientes de fibra óptica pagan, en la Figura 6.23 se muestra que las QNP estimadas para los diferentes grupos son relativamente similares; por lo que, a priori el monto que un cliente paga no es un factor que predice directamente el comportamiento de pago al considerar todos los perfiles de clientes.

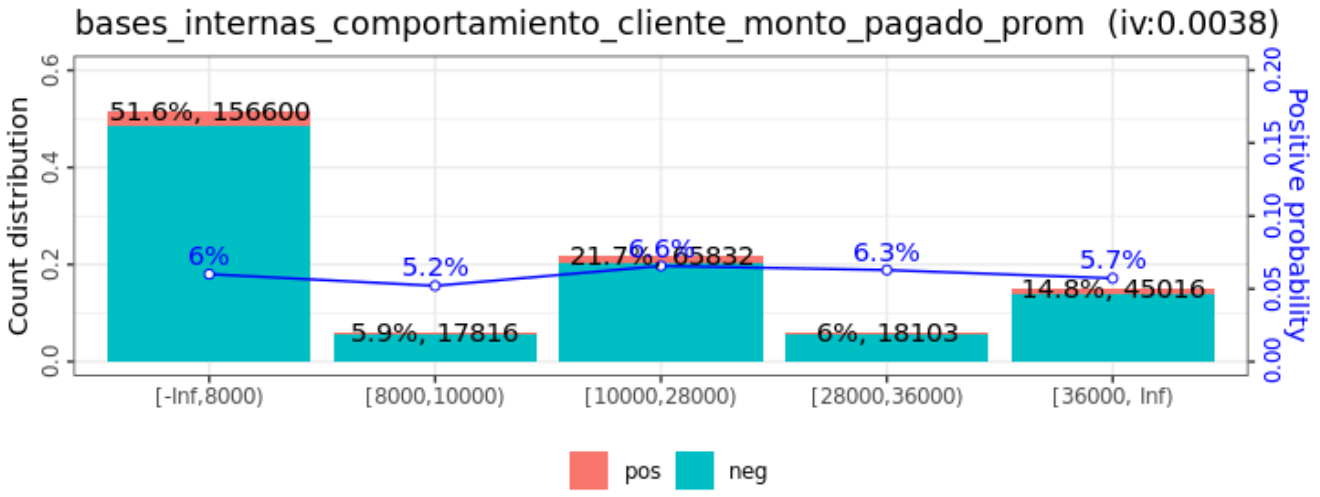


Figura 6.23: QNP agrupada para variable interna monto pagado promedio.

6.4. Modelamiento

Tal como se indica en los Alcances de este trabajo el foco de los modelos se encuentra en los servicios fibra óptica. En la tabla 6.1 se exponen los porcentajes de registros disponibles para las distintas antigüedades de los clientes, de acuerdo a esta tabla es posible inferir que la mayoría de los clientes que accedieron a la contratación del servicio fibra corresponden esencialmente a los segmentos de clientes antiguos y nuevos, por lo que los modelos para estos perfiles son de especial interés ya que concentran más del 95 % de las ventas realizadas del producto.

Tabla 6.1: Porcentaje de registros de acuerdo a la antigüedad de clientes de fibra óptica.

Tipo cliente	**Porcentaje de registros en fibra**
Nuevos	54.81 %
Semi nuevos	4.52 %
Antiguos	40.67 %

El flujo para construir los modelos es el mismo independiente del tipo de cliente (*véase* la Figura 6.24). En primer lugar se selecciona el segmento de clientes en fibra el cual puede corresponder a: nuevo, semi nuevo o antiguo. Una vez separados los registros correspondientes al segmento de interés se procede a trabajar la data sin transformación WoE; los algoritmos que se iteran consisten en regresión logística, random forest y gradient boosting machine, cada uno de ellos con diferentes técnicas de balanceo.

La utilización de diferentes métodos de balanceo es esencial ya que considerando todos los clientes que accedieron a fibra únicamente el 3.5 % corresponden a clientes con el comportamiento a predecir, es por este motivo que los modelos consideran escenarios: sin balanceo de clases, balanceo mediante sobremuestreo y submuestreo, y balanceo mediante la asignación de pesos a las categorías de la variable estudiada.

Para que los resultados de los modelos balanceados fuesen comparables las técnicas de balanceo utilizan los mismos parámetros: para el caso del balanceo mediante el sobre y sub muestreo (over & under sampling), la clase mayoritaria se redujo a la mitad mientras que la clase minoritaria se cuadriplicó; mientras que en el caso del balanceo por pesos de las clases se asignó un peso igual a 5 para los registros minoritarios y un peso igual a 1 para los correspondientes a la clase mayoritaria.

Una vez construidos todos los modelos previos, se procede a aplicar la transformación WoE a los datos y nuevamente iterar los mismos escenarios expuestos previamente esta vez con los valores de las variables agrupados. Es relevante destacar que el modelo final para cada escenario propuesto en el diagrama contempla el ajuste de los hiper parámetros de los algoritmos y la limitación de las variables más relevantes para cada modelo [1].

[1]La cantidad de variables puede ser como máximo 15 debido a que el sistema en el que se implementan los modelos posee esta limitación tecnológica.

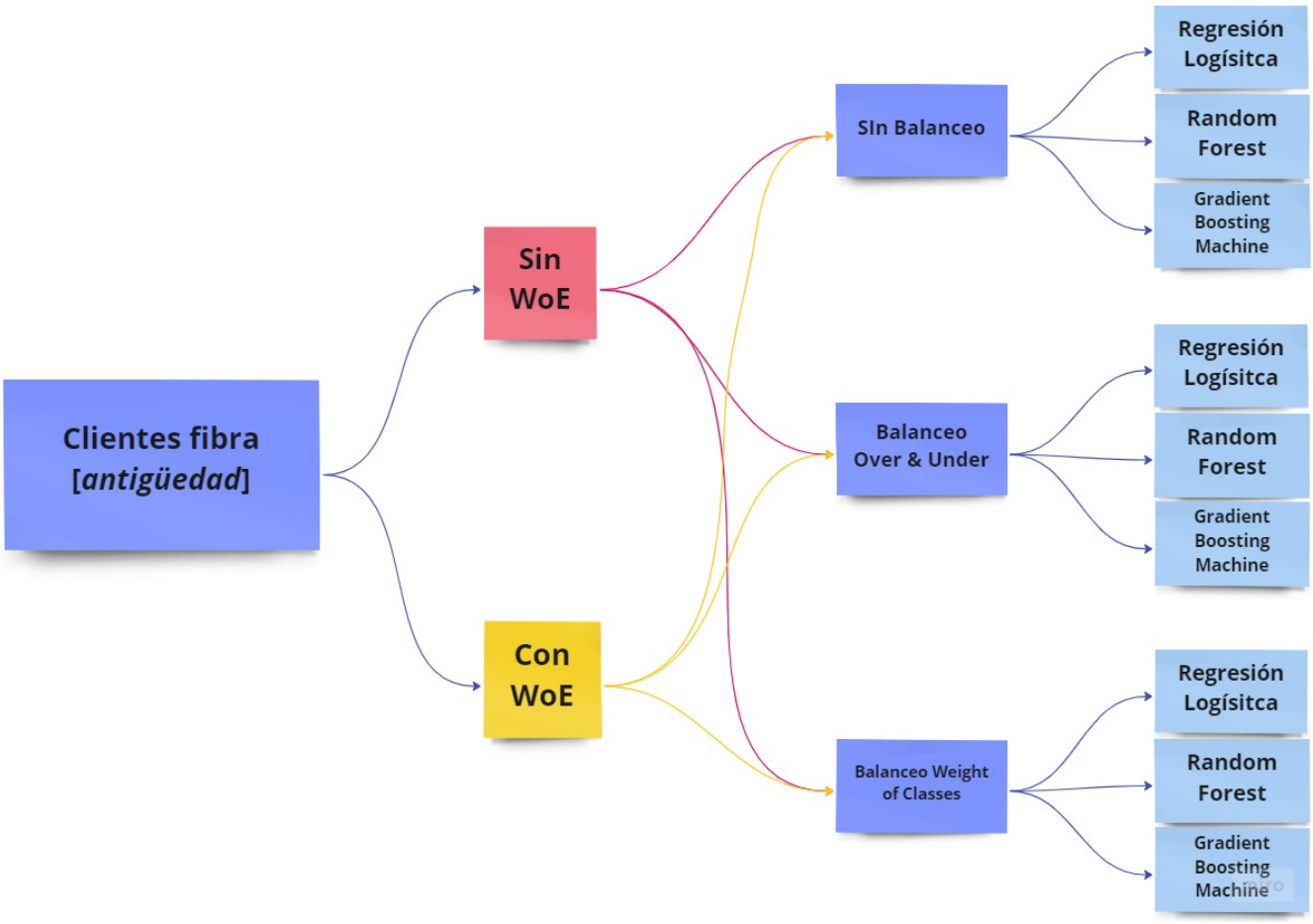


Figura 6.24: Esquematización del flujo de los distintos modelos elaborados.

La división de los datos para la base de entrenamiento y testeo queda definida por camadas de clientes, debido a que de esta forma es más intuitivo para el área de riesgo analizar los resultados. Para entrenamiento se utilizaron todos los registros comprendidos entre Agosto del 2020 a Diciembre de 2021, los cuales representan cerca del 73 % de la base; mientras que testeo abarca las camadas desde Enero 2022 a Mayo 2022, lo que representa un 27 % de los registros.

Las métricas utilizadas para la evaluación de modelos corresponden a 5 criterios: AUC, AUC Precision Recall, KS, PSI y el comportamiento de la curva Lift [2]; estos criterios son los utilizados por el equipo de riesgo para la toma de decisiones con respecto a los modelos operativos. Además, dado el segmento de clientes se exigen diferentes umbrales que cada modelo debiese cumplir dada la antigüedad del cliente con la compañía, estos criterios se enuncian en la Tabla 6.2.

[2]El significado y la interpretación de cada uno de estos criterios se encuentra en el Marco Conceptual.

Tabla 6.2: Criterios utilizados para la selección de modelos

Métrica	Antiguos	Nuevos y semi nuevos
AUC	$\geq$ 0,70	$\geq$ 0,65
AUC PR	$\geq$ 0,25	$\geq$ 0,15
KS	$\geq$ 0,35	$\geq$ 0,25
PSI	$\leq$ 0,10	$\leq$ 0,10
Lift	Decreciente	Decreciente

6.4.1. Clientes nuevos

El primer modelo en ser presentado pertenece al segmento de clientes **completamente nuevos**, este grupo incluye a todas las personas que en el sistema posean una antigüedad nula al momento de solicitar fibra óptica. Este perfil concentra cerca del 55 % de las ventas (*véase* Tabla 6.1), siendo la categoría con mayor volumen de registros.

En la Tabla 6.3 se presentan las principales métricas asociadas a los diferentes escenarios de iteración; ningún modelo construido para este perfil supera el umbral del KS definido en la Tabla 6.2, por lo que se presenta el modelo más cercano a ese objetivo. El algoritmo en cuestión corresponde a un **Random Forest con transformación WoE de sus datos y balanceado mediante la asignación de pesos**; la configuración que entrega el mejor desempeño consta de 36 árboles y una profundidad máxima igual a 6.

Tabla 6.3: Tabulación de métricas AUC, AUCPR y KS para todos los modelos nuevos fibra desarrollados.

Modelo			AUC	AUCPR	KS
Sin transformación WoE	Sin balanceo	Regresión Logística	0.6292	0.0800	0.1700
		Random Forest	0.6439	0.0852	0.1925
		Gradient Boosting Machine	0.6572	0.0924	0.2234
	Balanceo Over & Under sampling	Regresión Logística	0.6102	0.0580	0.1470
		Random Forest	0.6279	0.0612	0.1645
		Gradient Boosting Machine	0.6462	0.0774	0.1934
	Balanceo Weight of Classes	Regresión Logística	0.6278	0.2956	0.1639
		Random Forest	0.6478	0.3205	0.2005
		Gradient Boosting Machine	0.6453	0.3126	0.1908
Con transformación WoE	Sin balanceo	Regresión Logística	0.6560	0.0885	0.2243
		Random Forest	0.6584	0.0878	0.2312
		Gradient Boosting Machine	0.6553	0.0872	0.2223
	Balanceo Over & Under sampling	Regresión Logística	0.5992	0.0560	0.1450
		Random Forest	0.6219	0.0652	0.1735
		Gradient Boosting Machine	0.6452	0.0704	0.2094
	Balanceo Weight of Classes	Regresión Logística	0.6564	0.3170	0.2243
		Random Forest	0.6602	0.3214	0.2302
		Gradient Boosting Machine	0.6532	0.3159	0.2223

Las variables independientes incuidas en el modelo seleccionado se presentan en la Tabla 6.4, se

observa que todas estas poseen un origen externo lo cual se encuentra en concordancia con el perfil, ya que al ser clientes nuevos la empresa no cuenta con información acerca de su comportamiento de pago.

Tabla 6.4: Variables utilizadas en el modelo para el segmento de clientes nuevos.

Bases externas
Índice socio económico de emergencia
Bancos en que ha registrado movimientos
Avalúo fiscal bienes raíces
Tasación bien raíz
Cantidad documentos morosos
Monto documentos morosos
Cantidad de morosidades ult 36 meses
Monto de morosidades ult 36 meses
Cantidad total impagos
Monto total impagos
Cantidad impagos ult 36 meses
Antigüedad último documento moroso

Por motivos de confidencialidad no se presentan el peso exacto de cada variable, pero en la Figura 6.25 se presenta la importancia normalizada de cada una de ellas. De este gráfico es posible inferir que las variables externas que almacenan el ISE y la cantidad de bancos en los que el cliente potencial ha registrado movimientos son las más relevantes en la predicción.

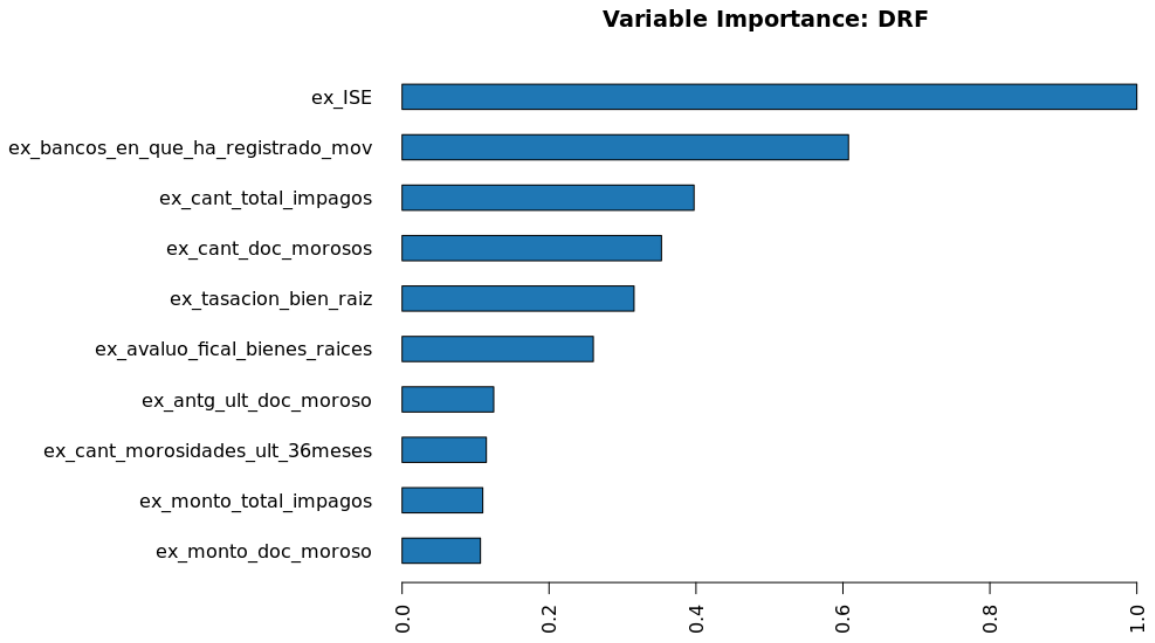


Figura 6.25: Importancia de las variables normalizada para el modelo nuevos fibra

La matriz de confusión de este modelo se expone en la Tabla 6.5, y las métricas principales construidas a partir de esta matriz se encuentran en la Tabla 6.6.

Tabla 6.5: Matriz de confusión para modelo antiguos fibra.

Observación/Predicción	**Positivo**	**Negativo**	**Total**
Positivo	3895	10634	14529
Negativo	1410	10590	12000
Total	5305	21224	**26529**

Tabla 6.6: Métricas de evaluación a partir de la matriz de confusión obtenida para el modelo antiguos fibra.

Precisión	**Sensibilidad**	**Exactitud**	**Especificidad**
0.7342	0.2681	0.5460	0.8825

Dado el desbalance de la data, se espera que la especificidad sea alta debido a que esta métrica determina los clientes que fueron correctamente designados como pagadores, es la tasa de verdaderos negativos identificados correctamente, a partir de este criterio es posible indicar que la clase mayoritaria esta siendo predicha con un 88 % de asertividad. Por su lado, la exactitud indica que el 55 % de los casos están siendo correctamente clasificados, pero dado que la especificidad es muy alta, el alto valor alcanzado en la exactitud se debe principalmente a la predicción correcta de la clase mayoritaria.

En base a lo expuesto recientemente tanto la precisión como la sensibilidad son métricas más interesantes de observar, ya que estas tienen el foco en la clase objetivo. La precisión alcanza el 73 % lo que indica que cerca de 3 de cada 4 clientes que el modelo predijo que no pagarían en efecto no lo hicieron; por otro lado sensibilidad señala que 1 de cada 4 clientes que no pagaron son etiquetados como no pagadores por el modelo.

En lo que respecta a las curvas ROC es posible observar que con la configuración indicada para este modelo entrega curvas de entrenamiento y testeo con comportamientos bastantes similares (*véase* Figura 6.26), además para la base testeo se alcanza un AUC del 0.6625.

En el gráfico de las curvas Lifts (*véase* Figura 6.27) no se observa sobre ajuste significativo en los primeros percentiles; además, tanto la curva de entrenamiento como la de testeo son estrictamente decrecientes. Finalmente, a partir de esta figura es posible indicar que la implementación de este modelo detecta 3.58 veces más casos positivos para el percentil 1 % en contraste a una selección aleatoria.

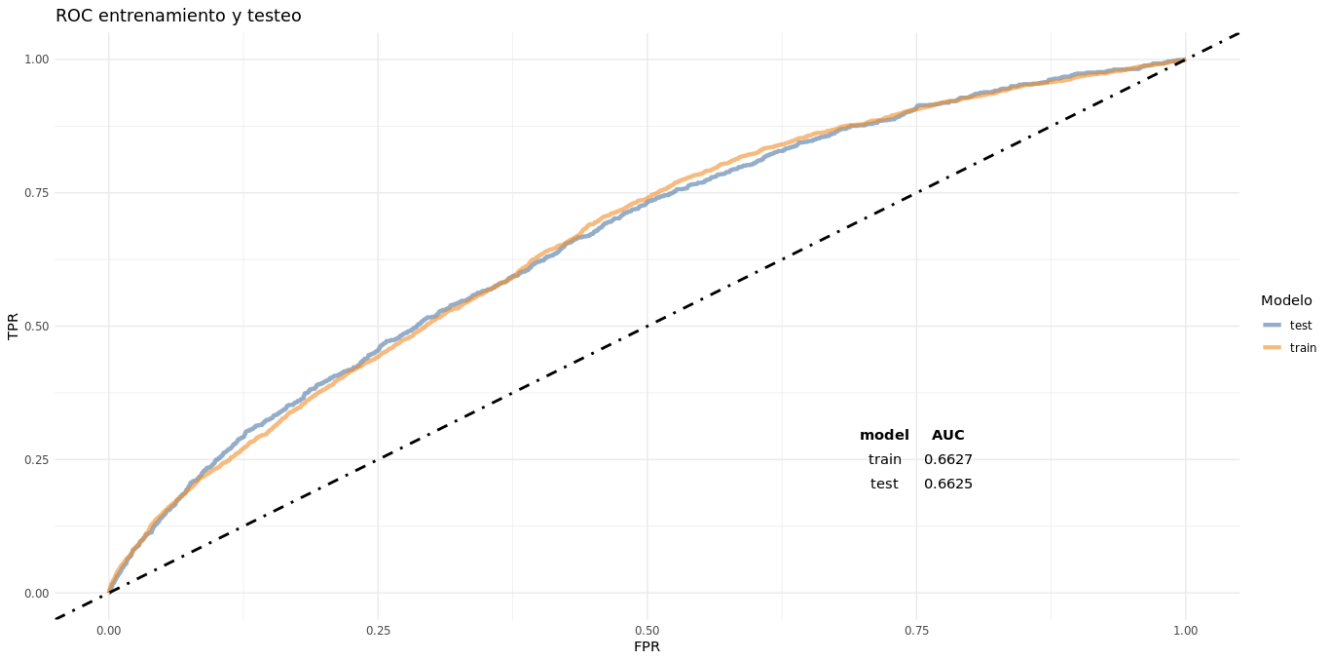


Figura 6.26: Curvas ROC y áreas bajo estas curvas para entrenamiento y testeo del modelo nuevos fibra.

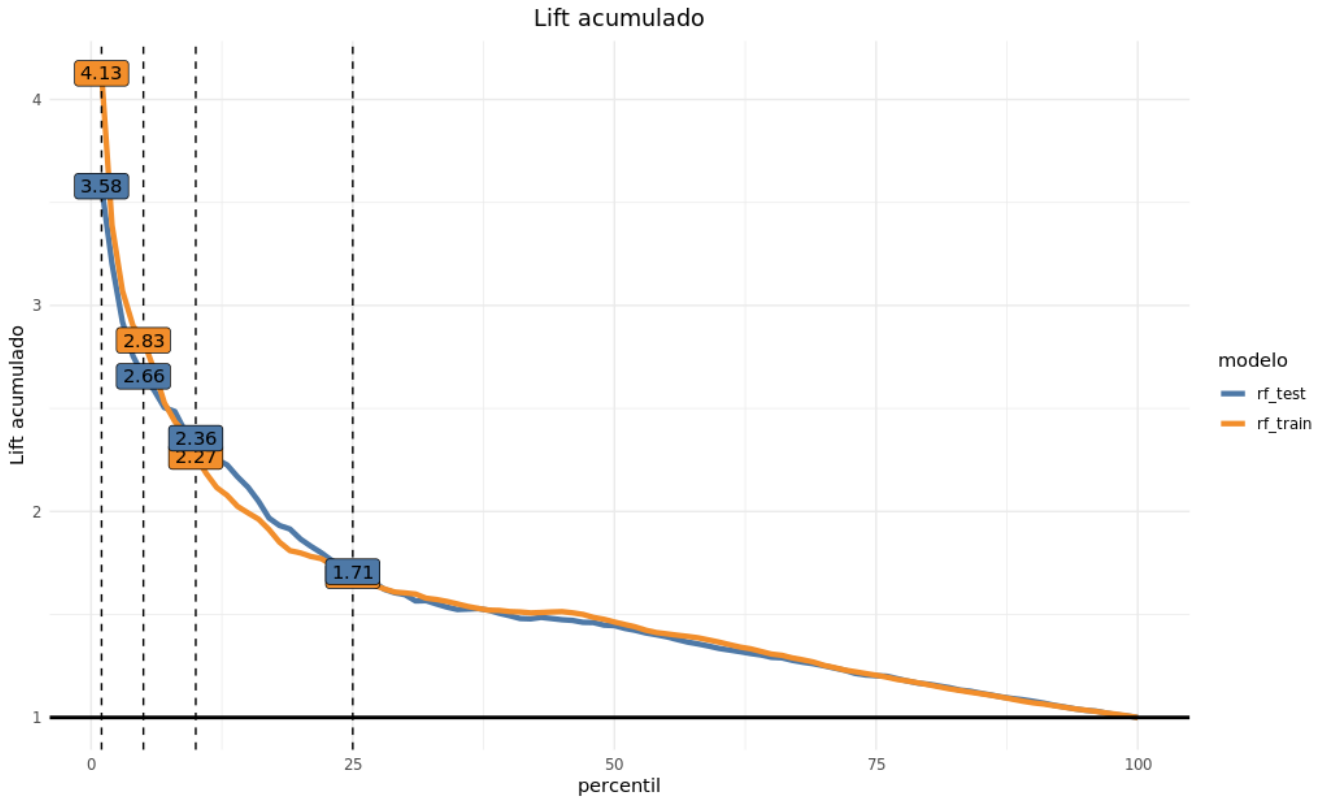


Figura 6.27: Comportamiento curvas Lift acumuladas de entrenamiento y testeo para modelo nuevos fibra, con percentiles 1 %, 5 %, 10 % y 25 % etiquetados.

Finalmente, las métricas de interés restantes se presentan en la Tabla 6.7. Se observa un AUCPR considerablemente menor que el AUC debido al desbalanceo de la data, sin embargo el trade-off entre sensibilidad y precisión se encuentra por sobre el umbral estipulado previamente. A su vez, el PSI bordea el 1 % esto indica que las variables independientes de testeo del modelo mantienen un comportamiento similar a las de entrenamiento, por lo que no debiese presentarse una desregularización del modelo que se origine debido a diferencias del comportamiento de las variables independientes. Por último, el KS evidencia que tan diferentes son la clase pagadora de la no pagadora, en este caso el modelo es capaz de diferenciar un mal pagador de un bueno con un 23.02 % de efectividad, dos puntos porcentuales menor al umbral establecido para este segmento.

Tabla 6.7: Criterios restantes utilizados en la evaluación y selección del modelo nuevos fibra.

Métrica	**AUCPR**	**KS**	**PSI**
Valor	0.3213	0.2302	0.0106

6.4.2. Clientes semi nuevos

A continuación se presentan las principales métricas obtenidas para los escenarios del segmento de clientes **semi nuevos**, esta clase engloba a todas las personas que al momento de solicitar su evaluación para la venta de fibra poseía una antigüedad con la compañía entre 1 y 5 meses. De acuerdo a la Tabla 6.1 esta porción de clientes solo alcanza el 4,5 % de las ventas, transformándola en la agrupación con menor cantidad de ventas para el servicio de fibra óptica.

Tabla 6.8: Tabulación de métricas AUC, AUCPR y KS para todos los modelos semi nuevos fibra desarrollados.

Modelo			**AUC**	**AUCPR**	**KS**
Sin transformación WoE	Sin balanceo	Regresión Logística	0.6668	0.0850	0.2546
		Random Forest	0.7378	0.1080	0.3494
		Gradient Boosting Machine	0.7272	0.0958	0.3495
	Balanceo Over & Under sampling	Regresión Logística	0.6388	0.3044	0.2316
		Random Forest	0.7108	0.4034	0.3394
		Gradient Boosting Machine	0.6972	0.3614	0.3204
	Balanceo Weight of Classes	Regresión Logística	0.6501	0.3304	0.2596
		Random Forest	0.7478	0.4254	0.3444
		Gradient Boosting Machine	0.7257	0.3744	0.3394
Con transformación WoE	Sin balanceo	Regresión Logística	0.7259	0.1081	0.3245
		Random Forest	0.7481	0.1180	0.3744
		Gradient Boosting Machine	0.7223	0.0983	0.3344
	Balanceo Over & Under sampling	Regresión Logística	0.6969	0.3686	0.3015
		Random Forest	0.7271	0.4093	0.3484
		Gradient Boosting Machine	0.6993	0.3103	0.3114
	Balanceo Weight of Classes	Regresión Logística	0.7257	0.3956	0.9844
		Random Forest	0.7612	0.4303	0.3993
		Gradient Boosting Machine	0.7211	0.3403	0.3594

En la tabulación de las métricas para los distintos escenarios (*véase* Tabla 6.8) es posible observar que para este segmento hay 11 modelos que alcanzan los umbrales definidos; sin embargo, fue

seleccionado aquel que entrega los valores de métricas más altos. El modelo en cuestión corresponde a un **Random Forest con transformación WoE de sus datos y balanceado mediante la asignación de pesos de clases**; este árbol obtiene su mejor desempeño al ser construido con una profundidad máxima de 4 y 40 árboles.

Las variables que emplea este modelo se presentan en la Tabla 6.9; de estas el 75 % corresponden a información de origen externo, situación que se encuentra en sintonía con la lógica del negocio, ya que dado que este tipo de clientes son relativamente nuevos la mayor cantidad de información para predecir debiese proceder de fuentes exteriores a la empresa. Con respecto a la variables más relevantes, se observa en la Figura 6.28 que para el caso de los clientes semi nuevos las variables con mayor importancia son: bancos en los que ha registrado movimientos, monto deuda castigada, índice socio económico de emergencia y la cantidad total de impagos que el potencial cliente tiene al momento de solicitar el servicio.

Tabla 6.9: Variables utilizadas en el modelo para el segmento de clientes semi nuevos, categorizadas de acuerdo al origen de su información.

Bases internas	**Bases externas**
Monto deuda condonada	Índice socio económico de emergencia
Número de facturas pagadas	Bancos en que ha registrado movimientos
Número de facturas	Tasación bien raíz
	Avalúo fiscal bienes raíces
	Cantidad documentos morosos
	Cantidad total impagos
	Antigüedad último documento moroso
	Cantidad de morosidades ult 36 meses
	Cantidad impagos ult 36 meses

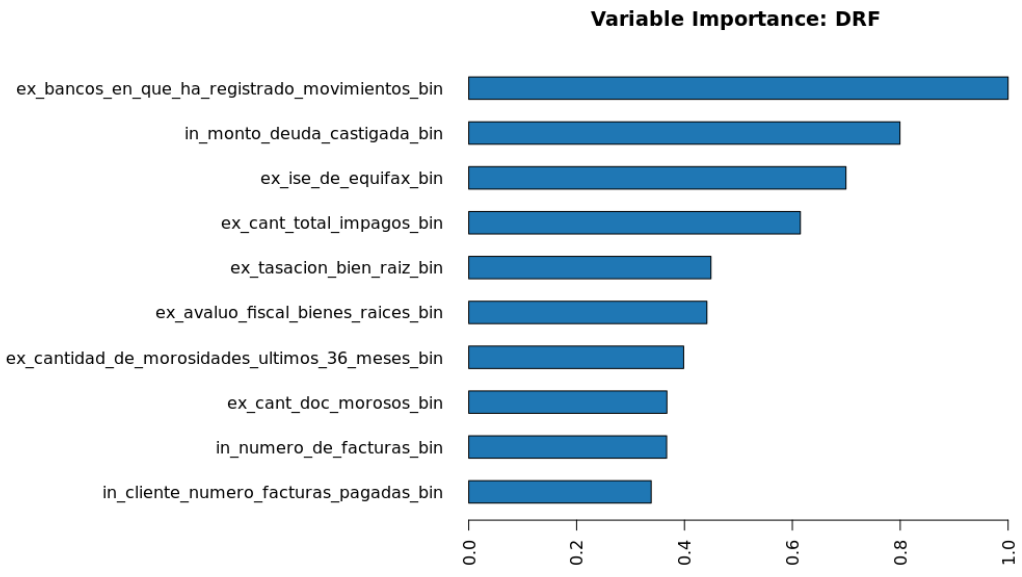


Figura 6.28: Importancia de las variables normalizada para el modelo semi nuevos fibra

La matriz de confusión del modelo se exhibe en la Tabla 6.10, y los indicadores que se construyen a partir de esta se presentan en la Tabla 6.11.

Tabla 6.10: Matriz de confusión para modelo semi nuevos fibra.

Observación/Predicción	**Positivo**	**Negativo**	**Total**
Positivo	300	391	691
Negativo	225	1784	2009
Total	525	2175	**2700**

Análogamente a los clientes nuevos se observa un alto valor de la especificidad (88,8 %) que se origina producto del desbalance de la data, dado que hay una mayor cantidad de registros de la clase pagadora el modelo predice a esta clase mejor que la no pagadora; esta situación también se ve reflejado en la exactitud la cual también tiene un valor elevado (77 %) producto de la facilidad para detectar a la clase pagadora.

La precisión alcanza un 57,1 %, lo que indica que el modelo clasifica correctamente a clase no pagadora cerca de 3 de cada 5 veces; mientras que la sensibilidad esta relatando que 2 de 5 clientes que no pagan son identificadores correctamente por el modelo.

Tabla 6.11: Métricas de evaluación a partir de la matriz de confusión obtenida para el modelo antiguos fibra.

Precisión	**Sensibilidad**	**Exactitud**	**Especificidad**
0.5714	0.4342	0.7719	0.8880

Las curvas ROC de entrenamiento y testeo presentadas en la Figura 6.29 evidencian comportamientos análogos, además el modelo en estudio alcanza un AUC del 0.7616 al momento de ser testeado. A su vez, las curvas Lifts para este segmento decrecen pero no en todo el conjunto de datos (*véase* Figura 6.30), por lo que estas curvas no son estricamente decrecientes; pese a esto, es posible indicar que para el percentil 1 % el modelo discrimna 4.72 veces mejor un no pagador que una selección aleatoria.

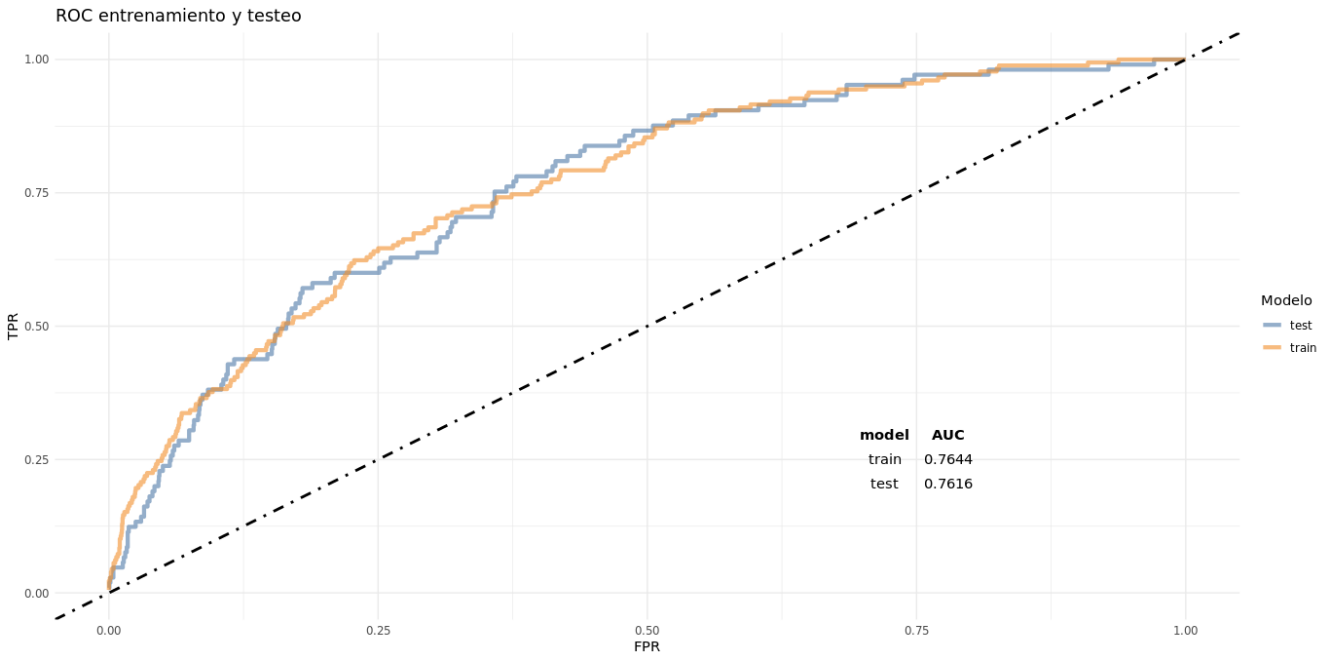


Figura 6.29: Curvas ROC y áreas bajo estas curvas para entrenamiento y testeo del modelo semi nuevos fibra.

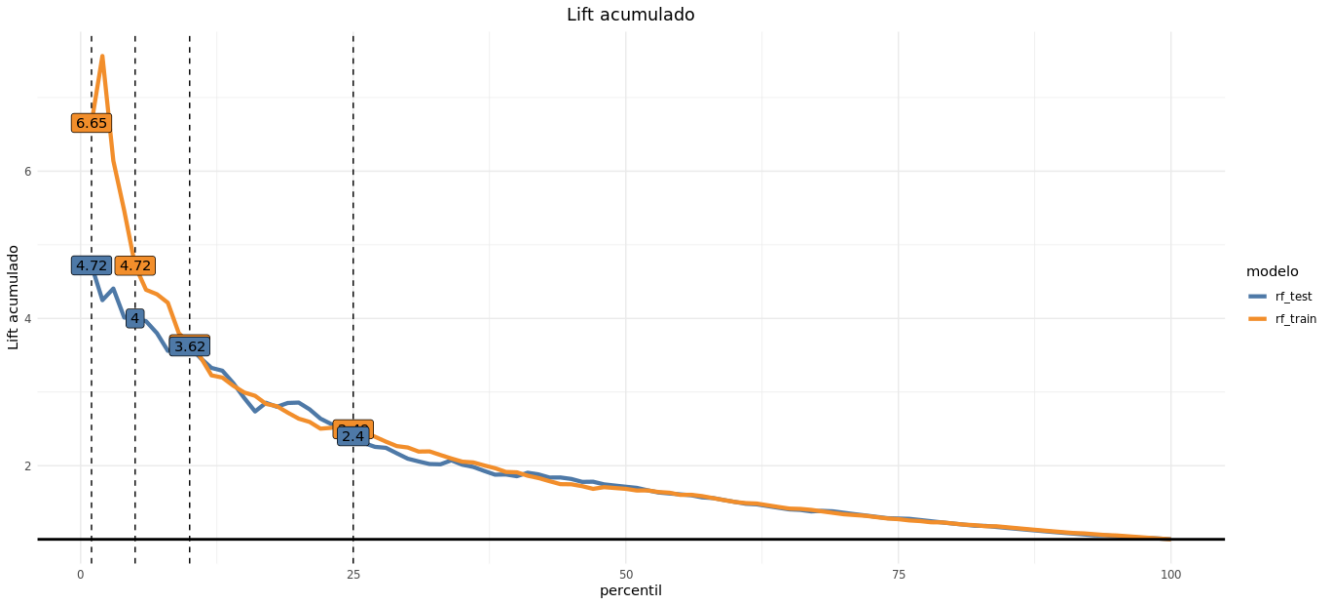


Figura 6.30: Comportamiento curvas Lift acumuladas de entrenamiento y testeo para modelo semi nuevos fibra, con percentiles 1 %, 5 %, 10 % y 25 % etiquetados.

En lo que respecta a la compensación entre precisión y sensibilidad es posible indicar que se encuentra por sobre el umbral establecido, ya que el AUCPR alcanza un valor igual a 0.4303 (**véase** Tabla 6.12), este valor como es de esperar es menor que el AUC ya que el foco esta de esta métrica se encuentra en la clase no pagadora. El KS por su lado nos narra que el modelo diferencia la clase no pagadora de la pagadora con cerca de un 40 % de asertividad.

Sin embargo, el PSI alcanza un 31,8 % (**véase** Tabla 6.12), este valor es considerablemente mayor al umbral establecido e indica que existe una alta propensión a que el modelo se desregularice producto de cambios en el comportamiento de las variables independientes; por lo que se recomienda revisar el modelo con el objeto de disminuir este riesgo potencial.

Tabla 6.12: Criterios restantes utilizados en la evaluación y selección del modelo semi nuevos fibra.

Métrica	**AUCPR**	**KS**	**PSI**
Valor	0.4303	0.3993	0.3177

6.4.3. Clientes antiguos

En función de lo expuesto al inicio de este capítulo se desarrollan todos los modelos para los diferentes escenarios esquematizados (*véase* Figura 6.24). En la Tabla 6.13 se presenta un resumen de las principales métricas de los modelos en cuestión, es relevante destacar que solo un modelo cumple todos los requisitos expuestos en la Tabla 6.2 para el segmento antiguos.

Tabla 6.13: Tabulación de métricas AUC, AUCPR y KS para todos los modelos antiguos fibra desarrollados.

Modelo			**AUC**	**AUCPR**	**KS**
Sin transformación WoE	Sin balanceo	Regresión Logística	0,6815	0,0543	0,2837
		Random Forest	0,7064	0,0689	0,3133
		Gradient Boosting Machine	0,6981	0,0807	0,2898
	Balanceo Over & Under sampling	Regresión Logística	0,6817	0,2160	0,2838
		Random Forest	0,7065	0,2610	0,3135
		Gradient Boosting Machine	0,6987	0,2773	0,2945
	Balanceo Weight of Classes	Regresión Logística	0,6743	0,2174	0,2532
		Random Forest	0,6995	0,2639	0,2945
		Gradient Boosting Machine	0,7299	0,2803	0,3557
Con transformación WoE	Sin balanceo	Regresión Logística	0,7174	0,0742	0,3275
		Random Forest	0,7182	0,0714	0,3308
		Gradient Boosting Machine	0,7148	0,0744	0,3275
	Balanceo Over & Under sampling	Regresión Logística	0,7180	0,2659	0,3281
		Random Forest	0,7188	0,2646	0,3317
		Gradient Boosting Machine	0,7154	0,2738	0,3285
	Balanceo Weight of Classes	Regresión Logística	0,7193	0,2688	0,3284
		Random Forest	0,7181	0,2662	0,3204
		Gradient Boosting Machine	0,7189	0,2752	0,3322

El modelo que se presenta a continuación es el seleccionado para el segmento de clientes **antiguos**, es decir para aquellas personas naturales que al momento de solicitar el servicio fibra posean una antigüedad con la compañía mayor o igual a 6 meses.

El algoritmo con mejor desempeño corresponde a un **Gradient Boosting Machine sin transformación WoE de sus datos y balanceado mediante la asignación de pesos**. El óptimo de esta configuración se alcanza con 40 árboles y un nivel de profundidad máximo igual a 4.

Las variables independientes utilizadas en este modelo se enuncian en la Tabla 6.14, de estas la mayoría (60 %) corresponden a variables de origen interno. Esta situación es deseable, debido a que significa que las variables internas construidas por la empresa se encuentran capturando un gran porcentaje del comportamiento estudiado. Las variables con mayor peso (*véase* Figura 6.31) son: monto morosidades, monto deuda castigada, facturas pagadas, monto promedio pagado y dias de mora interna.

Tabla 6.14: Variables utilizadas en el modelo para el segmento de clientes antiguos, categorizadas de acuerdo al origen de su información.

Bases internas	**Bases externas**
Número facturas pagadas	Monto morosidades mayor 36 meses
Monto deuda condonada	Bancos en que ha registrado movimientos
Días mora interna	Cantidad de impagos
Monto financiamiento	Antigüedad último protesto
Monto pagado promedio	
Antigüedad en meses	

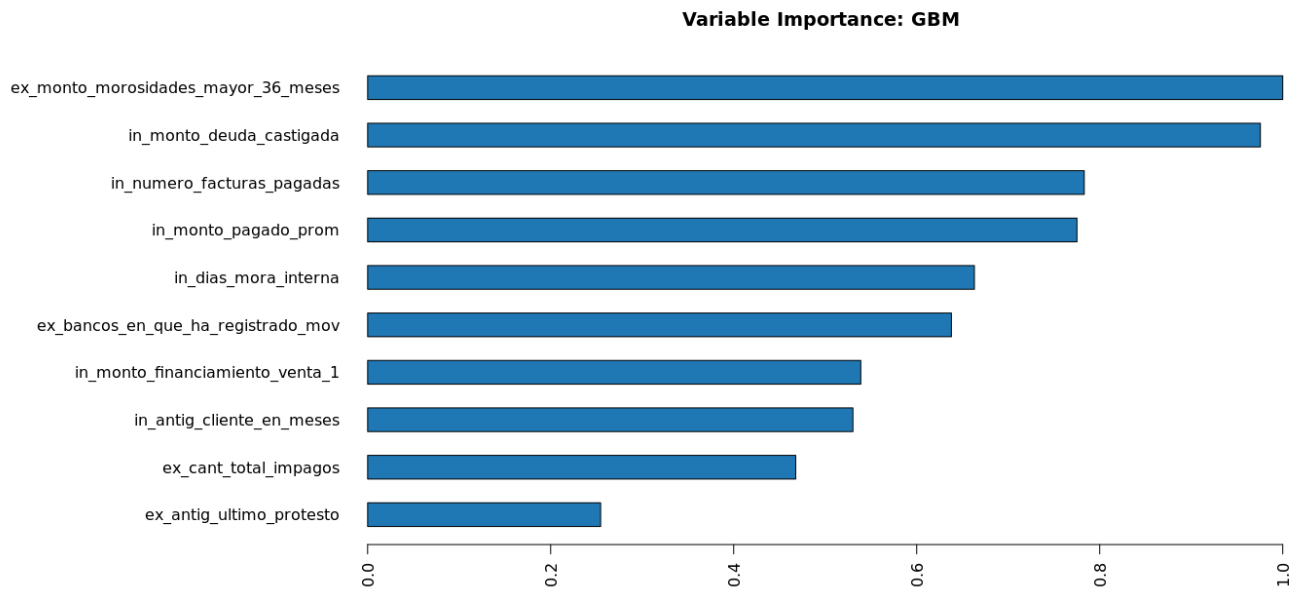


Figura 6.31: Importancia de las variables normalizada para el modelo antiguos fibra

La matriz de confusión entregada por el modelo se expone en la Tabla 6.15, a partir de esta se construyen las métricas habituales de desempeño las cuales se presentan en la Tabla 6.16. Nuevamente, debido al desbalance de la data se espera que la especificidad sea alta, en este caso esta métrica indica que un 93 % de los clientes pagadores son clasificados correctamente. A su vez, la exactitud evidencia que tan acertada es la predicción del modelo, en este caso el modelo predice acertadamente un 75 % de los casos ,-tanto pagadores como no pagadores-, pero dado que la especificidad es alta el gran valor alcanzado por la exactitud se debe principalmente a la clase mayoritaria.

Los otros indicadores son métricas más interesantes para estudiar debido a que su fin se encuentra en la clase de interés. La sensibilidad evidencia que tan bueno es el modelo identificando a los clientes no pagadores, en este caso el modelo logra etiquetar cerca de 1 de cada 4 no pagadores; mientras que la precisión expone que tan bien clasifica el modelo la clase no pagadora, por lo que cerca de 3 de cada 4 clientes que el modelo indico que no pagarían efectivamente no lo hicieron.

Tabla 6.15: Matriz de confusión para modelo antiguos fibra.

Observación/Predicción	**Positivo**	**Negativo**	**Total**
Positivo	1240	3721	4961
Negativo	940	13428	14368
Total	2180	17149	**19329**

Tabla 6.16: Métricas de evaluación a partir de la matriz de confusión obtenida para el modelo antiguos fibra.

Precisión	**Sensibilidad**	**Exactitud**	**Especificidad**
0.5688	0.2499	0.7586	0.9346

En lo que refiere a las curvas ROC de entrenamiento y testeo, -y sus respectivas áreas bajo las curvas-, es posible advertir que tienen comportamientos similares (*véase* Figura 6.32); alcanzando un AUC del 0.73 para la base de testeo.

En la Figura 6.33 se observa un leve sobre ajuste al inicio de las curvas lifts acumuladas; sin embargo, tanto para la base de entrenamiento como para la de testeo las curvas en cuestión son estrictamente decrecientes. Además, a partir de la misma figura es posible evidenciar que la producción de este modelo permite identificar 6.85 veces más casos de no pago para el percentil 1 % que una selección al azar.

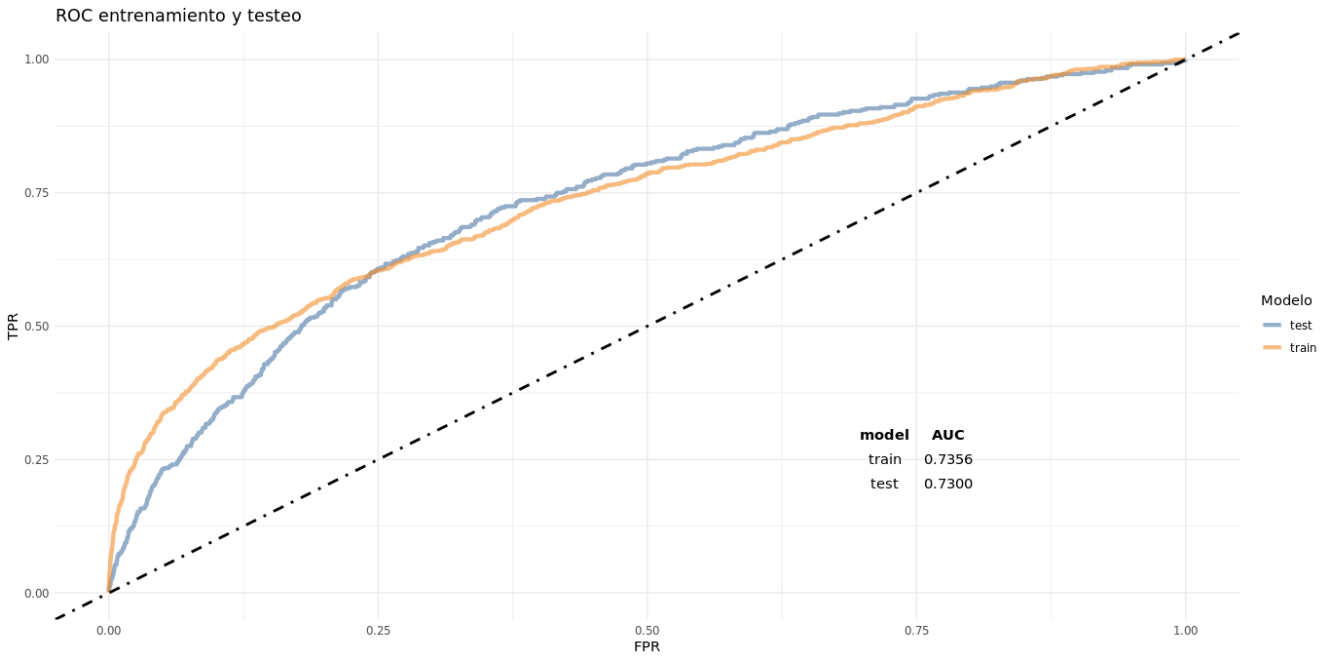


Figura 6.32: Curvas ROC y áreas bajo estas curvas para entrenamiento y testeo del modelo antiguos fibra.

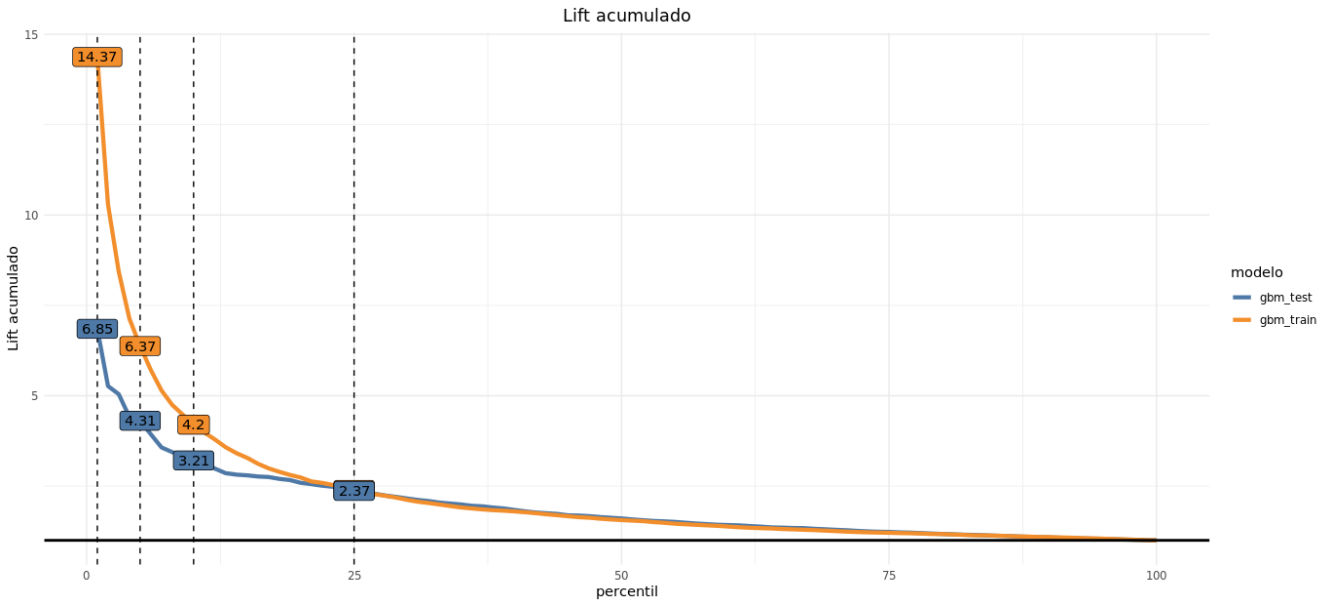


Figura 6.33: Comportamiento curvas Lift acumuladas de entrenamiento y testeo para modelo antiguos fibra, con percentiles 1 %, 5 %, 10 % y 25 % etiquetados.

El resto de los criterios prioritarios para la evaluación del modelo se presentan en la Tabla 6.17, de nuevo el AUCPR se encuentra por sobre el umbral establecido al inicio del capítulo (*véase* Tabla 6.2). Por su lado, el PSI rodea el 3,5 % lo que refleja que las variables independientes en la base de testeo poseen un comportamiento bastante similar a las mismas en las base de entrenamiento, por lo que existe poco riesgo de que el modelo sufra una desregurlación producto de una variación en

el comportamiento de estas variables. Finalmente, el KS muestra que tan diferentes son la clase pagadora de la no pagadora para el algoritmo, para este caso el modelo es capaz de diferenciar un mal pagador de uno bueno con un 35,57 % de efectividad.

Tabla 6.17: Criterios restantes utilizados en la evaluación y selección del modelo antiguos fibra.

Métrica	**AUCPR**	**KS**	**PSI**
Valor	0.2803	0.3557	0.035

6.5. Evaluación

A partir de la probabilidad de no pago pronosticada mediante el modelo, se construye un score de riesgo para cada uno de los clientes el cual oscila entre 0 y 1000 puntos. Las personas que obtienen una puntuación cercana a cero son clientes con una alta propensión al no pago (muy riesgosos); mientras que aquellos que obtiene un puntaje cercano a 1000 corresponden a clientes buenos pagadores (poco riesgosos).

Una vez scoreados los clientes se determinan las agrupaciones a utilizar en la estrategia de ventas, el único requisito que se impone para su construcción es que cada grupo tenga como mínimo el 5 % de los datos; luego, mediante WoE Binning se encuentran los cortes de score que permiten realizar agrupaciones de clientes con QNP similares. Para corroborar que esta segmentación es adecuada se realiza una prueba de significancia estadística[3], el objetivo que se persigue con esta evaluación es constatar que existen diferencias significativas entre los grupos construidos.

Luego de comprobar que la segmentación propuesta captura coherentemente las diferentes QNP, se presenta una estrategia de ventas la cual es evaluada financieramente; finalmente, dependiendo del beneficio financiero de la propuesta se concluye si es prudente de incorporar la estrategia.

Con el fin de resguardar la información en torno a los flujos financieros de la línea de negocio se presenta una simplificación de la valorización construida; por lo que, los valores de la Tabla 6.18 corresponden a una aproximación del comportamiento que son asumidos para el flujo. En la tabla en cuestión se presentan: valores promedios de la segunda factura aperturados por la antigüedad del cliente y el margen del negocio correspondiente al porcentaje de los ingresos que se perciben como utilidad al descontar los costos fijos y variables del negocio detallados adelante.

Tabla 6.18: Valores del negocio utilizados para la evaluación financiera

Variable	**Valor**
Precio prom 2da factura clientes nuevos y semi nuevos	25.000 CLP
Precio prom 2da factura clientes vigentes	28.000 CLP
Margen fibra sin considerar incobrable	11 %

Para determinar los **ingresos** se utiliza la relación 6.1.

$$Ingresos_{a,i,t} = Q_{a,i,t} \cdot P_a \tag{6.1}$$

Donde:

- $Ingresos_{a,i,t}$: Corresponde a los ingresos obtenidos dada una antigüedad a del cliente (nuevos, semi nuevos o vigentes), para una camada i de venta durante un mes t.
- $Q_{a,i,t}$: Corresponde a la cantidad de boletas facturadas para una camada i durante el periodo t, dada una antigüedad de los clientes a.

[3]Para más detalle véase Marco Conceptual.

- P_a : Corresponde al valor promedio de la segunda factura dada una antigüedad a del cliente (**véase** Tabla 6.18).

Por otro lado, los costos en los que se incurren durante el ejercicio de valorización son los siguientes:

- **Costos fijos:** corresponde esencialmente al costo de utilizar la **red** para entregar el servicio.
- **Costos variables:** corresponden a los montos destinados a la **captación, instalación y entrega** exclusiva del servicio contratado.
- **Incobrable:** Corresponde al monto de dinero que se debe provisionar para asumir la pérdida de los clientes que no pagan el servicio.

La aclaración previa es crucial ya que el monto disponible para cubrir el incobrable[4] no esta incluido en los costos fijos o variables; se construye de esta forma para simplificar la valorización, ya que al modificar las tasas de no pago el único tipo de costo que observa cambios en sus montos corresponde al incobrable.

Los **costos** quedan definidos por la relación 6.2.

$$Costos_{a,i,t} = CF_{a,i,t} + CV_{a,i,t} + I_{a,i,t} \tag{6.2}$$

Donde:

- $CF_{a,i,t}$**:** Costos fijos en que incurre una camada i en el periodo t, dada la antigüedad a de los clientes de la camada.
- $CV_{a,i,t}$**:** Costos variables en que incurre una camada i en el periodo t, dada la antigüedad a de los clientes de la camada.
- $I_{a,i,t}$**:** incobrable de una camada i en el periodo t, dada la antigüedad a de los clientes de la camada.

Finalmente, el incobrable queda definido por 6.3. , donde $TNP_{a,g}$ corresponde a la TNP 60 días posterior a la segunda factura dada una antigüedad a de los clientes y considerando los grupos g de la estrategia de ventas.

$$I_{a,i,t} = Ingresos_{a,i,t} \cdot TNP_{a,g} \tag{6.3}$$

Dado que las QNP y TNP obtiene valores bastante similares se asumirá que la TNP2F60 es igual a la QNP2F60, quedando una relación más sencilla para definir el incobrable por 6.4.

$$I_{a,i,t} = Ingresos_{a,i,t} \cdot QNP_{a,g} \tag{6.4}$$

[4]Incobrable corresponde la suma de dinero que no retorna, es la pérdida producto del no pago de un determinado servicio; se puede estudiar como una tasa del dinero no pagado por sobre el total recibido, o simplemente como la suma de los dineros perdidos. La utilidad de esta variable reside en que es un indicador que se construye en una amplia ventana de tiempo por lo que tiende a ser bastante acertado en el monto a provisionar, pero requiere observar a los clientes durante un largo periodo de tiempo.

Además, se procede a determinar el break even requerido para que el negocio sea capaz de cubrir sus requerimientos mínimos mediante la relación expuesta por A.Perez & J. Segundo [23]. De esta forma se obtuvo un break even promedio por camada, los cuales fueron nuevamente promediados para obtener el punto muerto representativo dada la antigüedad de un cliente, los valores obtenidos se presentan en la tabla 6.19.

$$BreakEven_{a,i,t} = \frac{CF_{a,i,t}}{P_a - \frac{CV_{a,i,t}}{Q_{a,i,t}}} \tag{6.5}$$

Tabla 6.19: Puntos muertos promedio dada la antigüedad de un cliente

Punto muerto	**Q clientes**
Nuevos	3222
Semi nuevos	266
Vigentes	2529

Finalmente, tal como se plantea en los objetivos de este trabajo el foco esencial de la estrategia esta en mantener los niveles de riesgo en rangos deseables, esto quiere decir que la estrategia en el escenario ideal permite disminuir las tasas de no pago aumentando las utilidades. Por lo que, no basta con reducir las tasas de no pago se debe observar si la cantidad de ventas es mayor al punto muerto y si las utilidades percibidas no caen drásticamente.

Adicionalmente, por el momento la estrategia de venta solo puede limitar la entrada de clientes, así que al momento de realizar la valorización se tiene como escenario base continuar la venta de todos los tramos y los escenarios de comparación corresponden al cierre de los tramos más riesgosos.

6.5.1. Clientes nuevos

Los tramos de riesgo propuestos para el perfil de clientes completamente nuevos se diagraman en la Figura 6.34; el grupo más riesgoso reúne los score que oscilan entre 0 y 670 puntos, posee una QNP del 12 % lo que indica que cerca de 1 de cada 8 clientes de este grupo no paga el servicio. Al mismo tiempo, el grupo menos riesgoso alcanza una QNP del 1.5 % y engloba las puntuaciones de 920 a 1000 puntos. Es relevante destacar, que en la Figura 6.34 se evidencia uno claro decrecimiento de la curva QNP a medida que se obtiene un mayor score.

Tabla 6.20: Desglose tramificación de clientes nuevos fibra.

Agrupación	**Score**	**Q total**	**Q positivo**	**Materialidad**	**QNP**
R1	[920,1000]	2334	36	10.5 %	1.5 %
R2	[8700,920)	3598	82	16.1 %	2.3 %
R3	[788,870)	5702	205	25.6 %	3.6 %
R4	[670,788)	9406	589	42.2 %	6.3 %
R5	[0 ,670)	1245	149	5.6 %	12 %

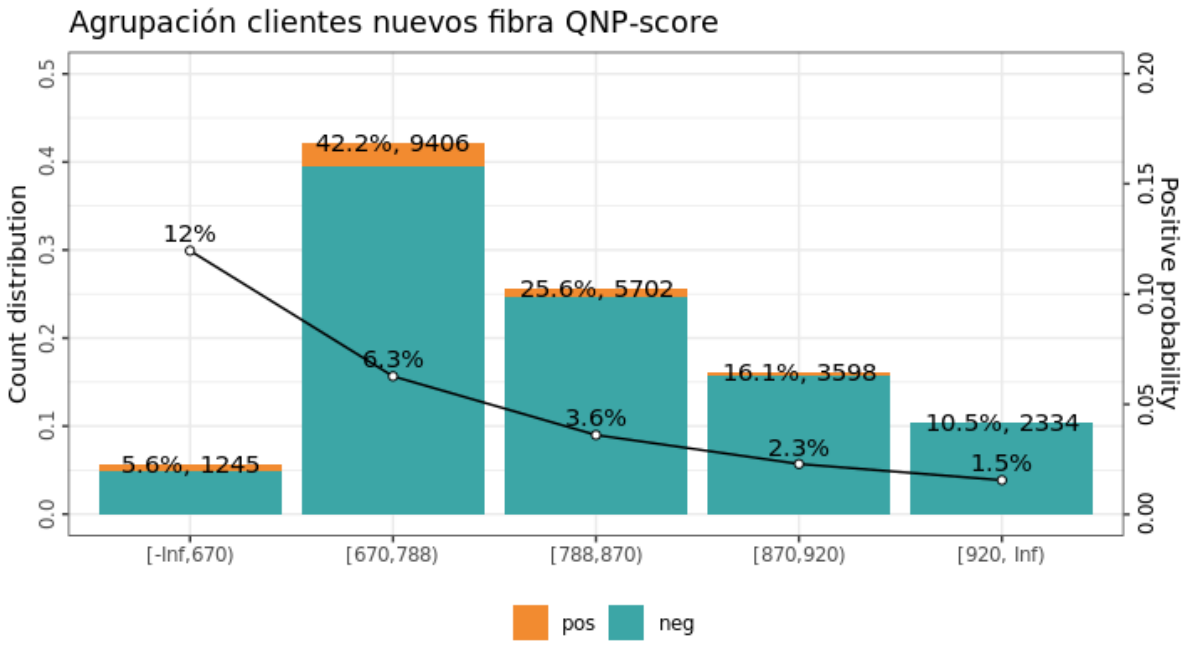


Figura 6.34: Tramificación construida para clientes completamente nuevos de fibra, a partir de los puntajes obtenidos

Los resultados del test de significancia estadística se presentan en la Tabla 6.21, a partir de este se puede corroborar que en efecto las tramificaciones construidas son lo suficiente diferentes, ya que el estadístico Z obtenido es mayor a 1,96 para todas las combinaciones de interés.

Tabla 6.21: Resultados test de significancia estadística al 95 % para tramos de riesgo clientes nuevos fibra. Sea R1 con menor nivel de riesgo, mientras que R5 agrupa a los clientes más riesgosos.

Hipótesis Nula	**Hipótesis Alternativa**	**Z**	**Resultado**
R1 = R2	R1 $\neq$ *R2*	10.29	Rechaza nula
R2 = R3	R2 $\neq$ *R3*	15.95	Rechaza nula
R3 = R4	R3 $\neq$ *R4*	26.32	Rechaza nula
R4 = R5	R4 $\neq$ *R5*	23.80	Rechaza nula

Siguiendo el camino detallado al inicio de este capítulo se construye la evaluación financiera para los clientes nuevos, en esta se evalúan las siguientes propuestas:

- **Mantener abierta la venta a todos los tramos:** es esencialmente el escenario actual ya que no se regula la admisión de clientes mediante su propensión al no pago de los servicios.
- **Cerrar la venta al tramo R5:** este tramo es el que posee la QNP más alta, en caso de aplicarse se cerraría la venta a aquellas personas que obtengan un score menor a 670 puntos.
- **Cerrar la venta al tramo R4 & R5:** estos tramos son los más riesgosos, e implica solamente vender fibra a aquellas personas que obtengan más 788 puntos de score.

Los principales resultados obtenidos de la evaluación se presentan en la tabla 6.22. En la Tabla 6.19 se indica que el punto muerto corresponde a 3222, por lo que de inmediato se rechaza la posibilidad de cerrar simultáneamente el tramo R4 y R5, ya que en caso de realizar esta acción no se alcanzan las ventas requeridas para cubrir el piso mínimo.

Por otro lado, en caso de cerrar únicamente el tramo R5 se aumentan las utilidades del año simulado en 18 millones; además, la QNP 60 días posterior a la segunda factura disminuye desde un 4.78 % a un 4.44 % (reducción del 8.94 %). Como es de esperar al disminuir la cantidad de clientes que no pagan (QNP) disminuye el monto adeudado y por ende la TNP2F60, pasa desde un 4.88 % a un 4.44 %.

Tabla 6.22: Resumen evaluación financiera de los diferentes escenarios de estrategia de ventas para segmento de clientes nuevos fibra.

Estrategia seguida	**Utilidad**	**Margen obtenido**	**Q ventas prom**	**QNP 2F60**	**TNP 2F60**
Mantener abierta la venta a todos los tramos	$ 2.510.011.374	6,52 %	4.138	4.78 %	4.88 %
Cerrar la venta al tramo R5	$ 2.528.212.491	6,96 %	3.907	4.35 %	4.44 %
Cerrar la venta al tramo R4 & R5	$ 1.720.900.714	8.57 %	2160	1.45 %	2.83 %

Basándonos en lo expuesto anteriormente, se propone que la mejor estrategia para los clientes nuevos corresponde en el cierre del tramo R5.

6.5.2. Clientes semi nuevos

La tramificación propuesta para el segmento de clientes semi nuevos se presenta en la Figura 6.35, de esta es posible inferir que las QNP son más altas que los otros perfiles de clientes, esta situación puede deberse a la menor cantidad de datos que contempla este segmento.

El grupo más riesgoso engloba a todas las personas que tienen un score estimado menor a 720 puntos, la QNP es 17,8 % lo que indica que más de 1 de cada 6 seis clientes de este segmento no cancela el servicio; mientras que en el segmento menos riesgoso menos de 1 de cada 100 no paga el servicio. Como era de esperar al igual que en la tramificación previa, se observa que la curva QNP decae a medida que se obtiene una mejor puntuación score.

Tabla 6.23: Desglose tramificación de clientes semi nuevos fibra.

Agrupación	**Score**	**Q total**	**Q positivo**	**Materialidad**	**QNP**
R1	[900,1000]	551	3	24.2 %	0.5 %
R2	[840,900)	658	14	28.9 %	2.1 %
R3	[800,840)	618	28	27.1 %	4.5 %
R4	[720,800)	335	39	14.7 %	11.6 %
R5	[0 ,720)	118	21	5.2 %	17.8 %

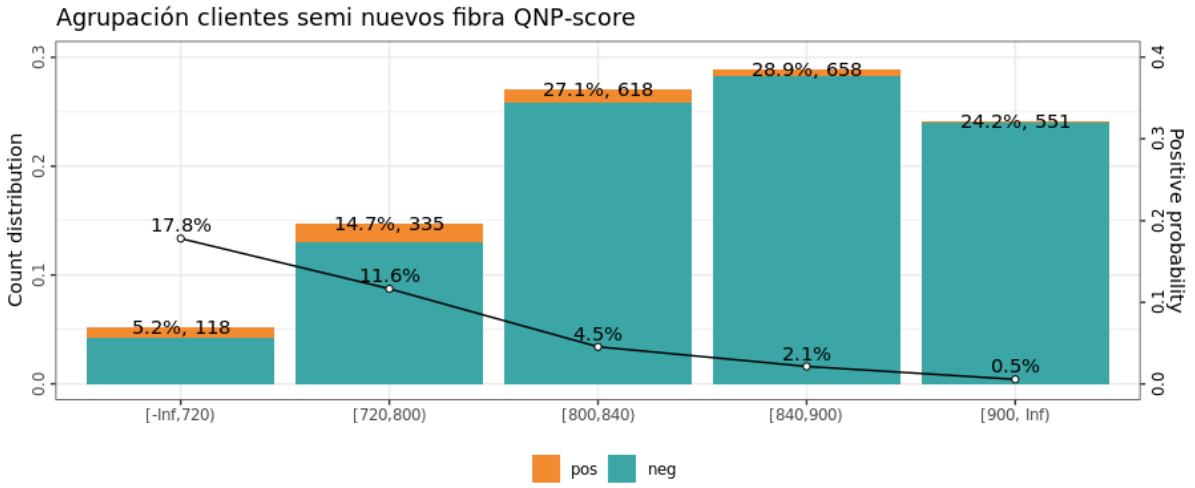


Figura 6.35: Tramificación construida para clientes semi nuevos fibra, a partir de los puntajes obtenidos

El resumen del test de significancia se presenta en la Tabla 6.24, de este es posible corroborar que en efecto la segmentación propuesta posee diferencias significativas entre grupos.

Tabla 6.24: Resultados test de significancia estadística al 95 % para tramos de riesgo clientes semi nuevos fibra. Sea R1 con menor nivel de riesgo, mientras que R5 agrupa a los clientes más riesgosos.

Hipótesis Nula	**Hipótesis Alternativa**	**Z**	**Resultado**
R1 = R2	R1 $\neq$ *R2*	3.93	Rechaza nula
R2 = R3	R2 $\neq$ *R3*	6.66	Rechaza nula
R3 = R4	R3 $\neq$ *R4*	9.32	Rechaza nula
R4 = R5	R4 $\neq$ *R5*	8.11	Rechaza nula

De igual manera que para el segmento de clientes nuevos, las estrategias de ventas propuestas consisten en el cierre de los tramo más riesgosos:

- **Mantener abierta la venta a todos los tramos:** escenario actual, no hay control sobre la admisión de los clientes a partir del cálculo de su propensión al no pago.
- **Cerrar la venta al tramo R5:** tramo con QNP más alta, en este escenario la acción consiste en privar la venta a los clientes con un score menor a 720 puntos y una antigüedad entre 1 y 6 meses.
- **Cerrar la venta al tramo R4 & R5:** tramos más riesgosos, implica impedir la venta a los clientes semi nuevos que tengan menos de 800 puntos de score de riesgo.

El resultado de la evaluación financiera de los diferentes escenarios es presentado en la Tabla 6.25. El punto muerto de este segmento de clientes corresponde a 266 ventas (*véase* Tabla 6.19), y dado que en todos los escenarios la cantidad de ventas promedio es mayor a ese valor no se excluye ninguno de los escenarios propuestos, ya que todos son capaces de entregar los ingresos necesarios para cubrir los costos esenciales.

Dado que todos los escenarios capturan ventas mayor al break even se debe escoger la alternativa que entregue la mayor utilidad, la cual consiste en cerrar la venta de los tramos R4 y R5, de esta forma el segmento captura 13 millones más que en caso de mantener la venta indiscriminada de fibra. Adicionalmente reduce la QNP2F60 de un 4.54 % a un 1.95 %, mientras que la TNP2F60 disminuye desde un 4.63 % a un 2.48 %; es relevante indicar que esta gran variación en las tasas en torno al no pago se debe a que la cantidad de ventas de este segmento son considerablemente menor que los tramos de clientes nuevos y vigentes, por lo que pequeñas variaciones se traducen en cambios importantes.

Tabla 6.25: Resumen evaluación financiera de los diferentes escenarios de estrategia de ventas para segmento de clientes semi nuevos fibra.

Estrategia seguida	**Utilidad**	**Margen obtenido**	**Q ventas prom**	**QNP 2F60**	**TNP 2F60**
Mantener abierta la venta a todos los tramos	$ 214.821.006	6,77 %	341	4.54 %	4.63 %
Cerrar la venta al tramo R5	$ 225.395.872	3.93 %	324	3.85 %	3.93 %
Cerrar la venta al tramo R4 & R5	$ 227.445.432	8.92 %	274	1.95 %	2.48 %

En base a todo lo expuesto previamente, la mejor estrategia a seguir para el segmento de clientes semi nuevos corresponde a limitar la venta solo a aquellas personas que obtengan un escore de riesgo mayor o igual a 800 puntos.

6.5.3. Clientes antiguos

Para los clientes antiguos la tramificación de riesgo propuesta es la que se presenta en la Figura 6.36, de esta es posible señalar que los clientes que obtengan un score entre 0 y 780 puntos poseen la QNP más alta de la cartera (10,7 %), donde 1 de cada 10 clientes no paga el servicio prestado; mientras que para el tramo menos riesgoso, -de 950 a 1000 puntos-, menos de 1 de cada 100 personas no paga el servicio. Como es de esperar la cantidad de personas que no pagan va disminuyendo a medida que aumenta la puntuación.

Tabla 6.26: Desglose tramificación de clientes antiguos fibra

Agrupación	**Score**	**Q total**	**Q positivo**	**Materialidad**	**QNP**
R1	[950,1000]	3943	28	22.4 %	0.7 %
R2	[930,950)	4907	58	27.8 %	1.2 %
R3	[905,930)	4307	88	24.4 %	2.0 %
R4	[780,905)	3569	166	20.2 %	4.7 %
R5	[0 ,780)	899	96	5.1 %	10.7 %

Mediante el test de significancia estadística (*véase* Tabla 6.27), se corrobora que la tramificación propuesta agrupa conjuntos estadísticamente diferentes.

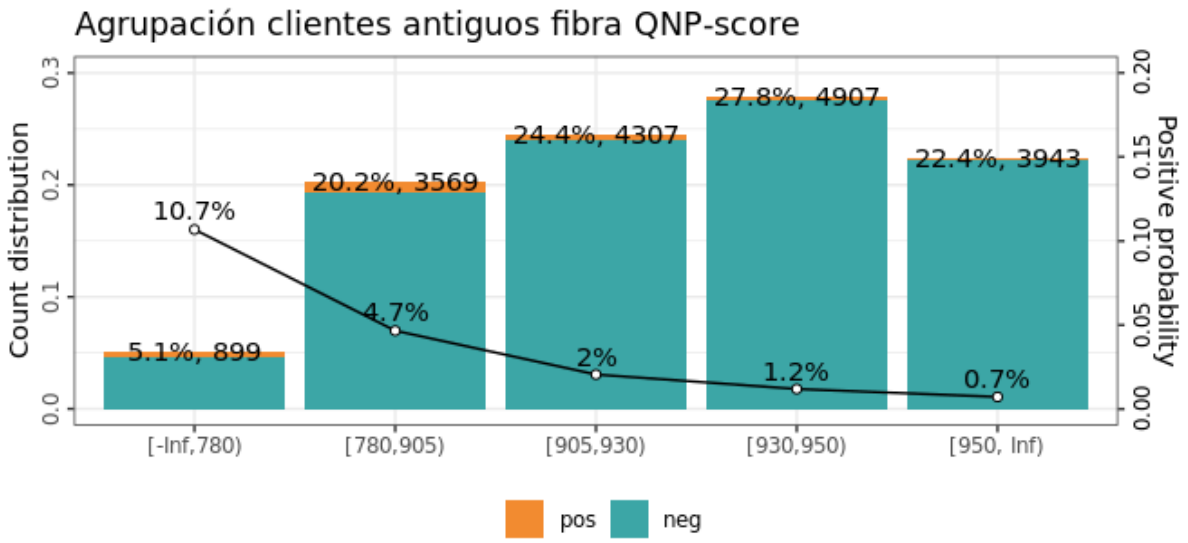


Figura 6.36: Tramificación construida para clientes antiguos fibra, a partir de los puntajes obtenidos

Tabla 6.27: Resultados test de significancia estadística al 95 % para tramos de riesgo clientes antiguos fibra. Sea R1 con menor nivel de riesgo, mientras que R5 agrupa a los clientes más riesgosos.

Hipótesis Nula	**Hipótesis Alternativa**	**Z**	**Resultado**
R1 = R2	R1 $\neq$ *R*2	9.06	Rechaza nula
R2 = R3	R2 $\neq$ *R*3	12.27	Rechaza nula
R3 = R4	R3 $\neq$ *R*4	16.75	Rechaza nula
R4 = R5	R4 $\neq$ *R*5	17.56	Rechaza nula

De manera análoga a los segmentos previos se proponen los siguientes escenarios a evaluar financieramente:

- **Mantener abierta la venta a todos los tramos:** situación actual, no hay control sobre la admisión de los clientes a partir del cálculo de su propensión al no pago.
- **Cerrar la venta al tramo R5:** este es el tramo más riesgoso (posee QNP más alta), la acción propuesta consiste en cerrar la venta a los clientes antiguos que obtengan un score menor a 780 puntos.
- **Cerrar la venta al tramo R4 & R5:** estos tramos son los mas riesgosos, la acción propuesta en este escenario consiste en impedir la venta de fibra a aquellos clientes vigentes que obtengan un score de riesgo menor a 905 puntos.

La síntesis del resultado de la evaluación financiera es presentada en la Tabla 6.25. Desde la Tabla 6.19 se infiere que el punto muerto de este segmento de clientes corresponde a 2529 ventas, y dado que al cerrar el conjunto a los tramos R4 y R5 no se alcanza ese umbral este escenario es desechado ya que no entrega los ingresos mínimos requeridos.

Al cerrar el tramo más riesgoso (únicamente R5), se observa una disminución de la utilidad obtenida de 2 millones durante la valoración anual construida. Sin embargo, permite movilizar la QNP 60 días posterior a la segunda factura desde un 2.48 % a un 2.03 % (reducción del 17.84 %), mientras que la TNP2F60 se moviliza desde un 2.52 % hasta un 2.07 %.

Tabla 6.28: Resumen evaluación financiera de los diferentes escenarios de estrategia de ventas para segmento de clientes antiguos de fibra.

Estrategia seguida	**Utilidad**	**Margen obtenido**	**Q ventas prom**	**QNP 2F60**	**TNP 2F60**
Mantener abierta la venta a todos los tramos	$ 2.787.723.304	8.48 %	3.155	2.48 %	2.52 %
Cerrar la venta al tramo R5	$ 2.785.638.921	8.93 %	2.994	2.03 %	2.07 %
Cerrar la venta al tramo R4 & R5	$ 2.371.508.302	9.67 %	2.355	0.98 %	1.33 %

Para este segmento las tasas de no pago son bajas en contraste a los segmentos nuevos y semi nuevos, y dado que la mayor utilidad la entrega el escenario actual se propone mantener abierta la venta indiscriminada a todos los clientes antiguos de la compañía.

Dado que la construcción de los modelos tiene por objeto mejorar la discriminación de los clientes que ingresan a la compañía, se realiza un estrés de las QNP para conocer el punto desde el cual el cierre del tramo R5 comienza a percibir una utilidad mayor al escenario actual. Se obtuvo que el cierre de R5 es rentable desde que la QNP de este tramo comienza a tener un valor mayor al 12,3 %.

6.6. Despliegue

Para lograr que la implementación sea exitosa se requiere de un mayor nivel de intervención del área de Riesgo; posterior al desarrollo del modelo y la construcción de su respectiva estrategia se presentan los resultados obtenidos al área interesada (*véase* Figura 6.37). El objetivo de esta presentación es que Riesgo apruebe la estrategia construida; no obstante, habitualmente se solicitan algunas modificaciones por lo que suele tomar una par de sesiones lograr un consenso.

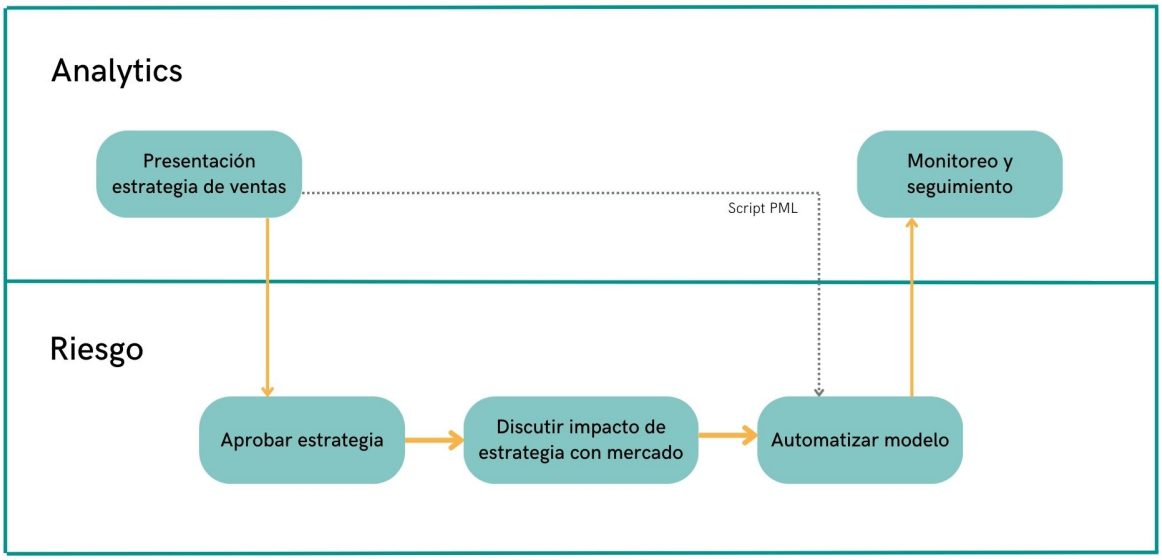


Figura 6.37: Diagrama de responsables y sus actividades asociadas en la fase de despliegue.

Una vez aprobada la estrategia por el Área de Riesgo, estos asumen la responsabilidad de notificar y discutir el impacto potencial de su implementación con el Área Comercial.

Toda la comunidad debe estar de acuerdo con la estrategia para comenzar con la fase de automatización del modelo. Teóricamente, Riesgo continúa siendo el responsable en esta etapa debido a que ellos poseen el acceso al sistema y cuentan con un equipo especializado en el manejo de la infraestructura donde son ejecutados los modelos; sin embargo, dado que todos los modelos a implementar son árboles, -y que el equipo de automatización se especializa en regresiones-, Analytics ha asumido el rol de parametrizar los modelos mediante la librería *r2pmml* la cual es compatible con la infraestructura.

Además de la conformación del árbol, la construcción de este script debe replicar todo el preprocesamiento construido previo a la ejecución del modelo. En esta etapa suelen surgir una serie de inconvenientes, entre los más recurrentes se tiene:

- **Versión r2pmml** debe encontrarse dentro de las compatibles con la infraestructura.

- El árbol debe ser construido con la **notación de variables del sistema**, por lo que los nombres de estas deben ser modificados de acuerdo al diccionario disponible.
- Se recomienda iterar el modelo en diferentes **escenarios bordes** de las variables, ocasionalmente al implementar los modelos estos suelen arrojar valores levemente diferentes a los esperados producto de las discrepancias en estos casos.
- Dejar claramente estipulado el **tratamiento de los valores NA**, estos generalmente son: reemplazados por mediana, promedio o cero, también suelen ser categorizados.

  Para el modelo de clientes nuevos y semi nuevos, ambos Random Forest con transformación WoE de la data, los missing values se tratan como una categoría más. Sin embargo, para el árbol de los clientes antiguos, un GBM sin transformación WoE, los valores faltantes son reemplazados por el valor promedio de la variable.
- En caso de que el problema se origine producto de errores en los valores almacenados en la variable, comunicar esta situación a Riesgo para que el equipo se comunique con el bureau financiero que entrega la información.

Es relevante indicar que la implementación del modelo se realiza a determinados dígitos verificadores, con el fin de mantener un muestra de control y otra de testeo.

Finalmente, la última etapa consiste en el monitoreo y seguimiento del modelo, de la cual Analytics es responsable (*véase* Figura 6.37). Para esta fase se construyen scripts mediante los cuales se estudia el comportamiento de la cantidad de ventas, QNP y TNP para las respectivas camadas.

Para las primeras camadas no es posible determinar la QNP ni TNP producto del desfase entre la camada y el comportamiento estudiado (60 días posterior a al segunda factura). Motivo por el cual, para aquellos primeros grupos de clientes solo se puede monitorear la cantidad de ventas para los dígitos verificadores en los que es implementado el modelo, se espera que no exista una diferencia significativa entre las ventas del grupo de control y el de testeo.

Dos meses posterior a la primera camada se puede determinar el primer acercamiento a los clientes no pagadores mediante el cálculo de la QNP y TNP 30 días posterior a la primera factura. Estas tasas serán numéricamente mayores que la QNP y TNP 60 días posterior a la segunda factura, pero si al graficar la QNP en los diferentes tramos de score de la estragia se observa que la curva posee un comportamiento decreciente a medida que aumenta el score, es un indicador de que el modelo esta categorizando correctamente a los clientes.

Capítulo 7

Conclusiones

7.1. Conclusiones generales

La venta de un servicio en el mercado no está exenta de incertidumbre, es por este motivo que existe una amplia investigación para hacer frente al trade off entre ventas y riesgo. La motivación de este trabajo se sustenta en aquel principio: **vender lo máximo posible sin percibir un aumento considerable en los niveles de riesgo**.

Debido a un creciente aumento de la demanda del servicio de fibra óptica en Chile, la empresa en señal de respuesta ha decidido invertir $106 millones de dólares para robustecer la infraestructura de esta tecnología y así ampliar su cobertura; sin embargo, este importante crecimiento ha conllevado que los indicadores de riesgo aumenten considerablemente durante el 2020 y 2021.

Esta situación vuelve prioritaria la estimación de los niveles de riesgo de fibra óptica. Para lograr predecir el comportamiento de los indicadores de interés se construyen modelos de clasificación binaria que buscan capturar el comportamiento de pago de los potenciales clientes de fibra óptica que intentan acceder a algún producto de la línea. De esta forma, mediante la predicción de la probabilidad de no pago, es posible proyectar las futuras tasas de no pago del servicio y así construir estrategias que permitan movilizar las tasas de riesgo hacia rangos deseables.

La comprensión del comportamiento de los clientes potenciales, requiere una investigación exhaustiva del funcionamiento del negocio y los actuales clientes. Dentro de los primeros hallazgos se encuentra la segmentación de los clientes de fibra óptica de acuerdo a su antigüedad: se consideran clientes **nuevos** a aquellos que nunca han accedido a un servicio de la empresa, **semi nuevos** los que han accedido a otro tipo de productos recientemente y **antiguos** serán aquellos que posean otro producto o servicio de la empresa desde hace más de seis meses. Es relevante destacar que los clientes nuevos y antiguos concentran el 95 % de los casos.

Para entrenar la propensión al no pago se trabaja con una variable objetivo binaria la cual toma una valor igual a 1 en caso de que el cliente adeude un saldo mayor a cero 60 días posterior al recibimiento de la segunda factura del servicio, y 0 en caso de que esté al día con el pago del servicio.

El segmento de clientes **nuevos** concentra el mayor volumen de ventas (55 %), pero también corresponde a uno de los perfiles con las QNP y TNP más altas. Para la predicción del comportamiento de los potenciales clientes con esta antigüedad, se selecciona un random forest que incluye 12 variables externas, entre las cuales se destaca el índice socio económico de emergencia y la cantidad de bancos en los que el solicitante ha registrado movimientos.

A partir de las predicciones del no pago, se construye una estrategia de ventas que permite segmentar a los clientes en base a su riesgo, la recomendación para este segmento consiste en **cerrar la venta del tramo más riesgoso (R5)**, por lo que todos aquellos clientes nuevos que obtengan un score de riesgo menor a 670 no debiesen acceder a la venta de fibra. De esta manera se aumentan las utilidades en 18 millones; además, se reduce la QNP2F60 de este segmento en un 8.94 % y la TNP2F60 en un un 9 %.

En cuanto al segmento de clientes **semi nuevos**, se observa una concentración del 5 % de los casos y un comportamiento bastante similar de la QNP y TNP de los clientes nuevos. Para la predicción del comportamiento de las personas que posean esta antigüedad al momento de solicitar el servicio fibra se construye un random forest que posee 3 variables internas y 9 variables externas, dentro de las variables con mayor relevancia se destaca: la cantidad de bancos en los que el solicitante ha registrado movimientos, la existencia de un monto adeudado con la empresa anteriormente, el índice socio económico de emergencia y la cantidad de impagos con otras empresas al momento de realizar la solicitud.

A través de la predicción del no pago, se construyen 5 segmentos para clasificar el riesgo de las personas. Luego de evaluar diferentes escenarios, se concluye que la estrategia de ventas para este perfil consiste en el **cierre de los segmentos más riesgosos (R4 & R5)**, esto implica que no debe concretarse la venta de fibra óptica a las personas que de este segmento obtengan un score de riesgo entre 0 y 800 puntos. La implementación de esta estrategia permite capturar 13 millones más de utilidades; además de reducir la QNP 60 días posterior a la 2da factura un 15.2 % y la TNP 2F60 en un 15.1 %.

Finalmente, se tiene al segmento de clientes **antiguos** el cual concentra cerca del 40 % de los casos; este perfil de cliente posee tasas de riesgo con desempeños bastante mejores que los segmentos previamente expuestos. La predicción del comportamiento de los clientes de este perfil se realiza mediante un gradient boosting machine que cuenta con 10 variables, siendo 6 de ellas de origen interno. Dentro del árbol las variables que más importancia tienen son: monto morosidades, monto adeuado con la empresa previamente, cantidad de facturas pagadas y el monto promedio que suele pagar el cliente.

Mediante la propensión al no pago predicha, se construye una nueva estrategia de ventas que cuenta con 5 segmentos de clientes. Posterior a la evaluación de diferentes escenarios con la tramificación propuesta es posible concluir que por el momento no es recomendable cerrar la venta

a ningún grupo de este perfil de clientes. Este resultado se debe esencialmente al buen comportamiento de pago de este tipo de clientes, ya que poseen una QNP2F60 de un 2.48 % mientras que la TNP2F60 bordea el 2.52 %; sin embargo, resultado estresar la valorización es posible concluir que el cierre del tramo más riesgoso es económicamente rentable si la QNP de R5 alcanza un valor mayor al 12,3 % (actualmente posee un 10.7 %).

7.2. Recomendaciones y trabajo futuro

Debido a los alcances de este trabajo solo se logra desarrollar modelos para el servicio fibra óptica, por lo que la primera recomendación es iterar la misma metodología para los servicios inalámbricos y de TV light de la línea de negocio. Pese a que estratégicamente se busca disminuir la venta de estos servicios, es inevitable que esa transición conllevará un periodo donde se deberá seguir monitoreando los niveles de riesgo de estos servicios, e idealmente se debe contar con información suficiente para tomar acciones en caso de que el riesgo continué aumentando en estos productos.

Por otra parte, es deseable robustecer la estrategia de ventas, por el momento solo se puede decidir si vender o no a un determinado grupo de clientes pero a futuro es posible construir una estrategia que además indique que plan podría ser el óptimo para un determinado cliente potencial.

Finalmente, la predicción del comportamiento de no pago de un cliente puede ser de utilidad para otras iniciativas; por ejemplo, puede utilizarse la probabilidad de no pago como input para otros modelos de predicción del comportamiento, como un modelo de gestión de cobranzas o perdonazos.

Bibliografía

[1] Entel. (s/f). Nuestra compañía. Revisado en Abril 2022, del sitio web: https://informacioncorporativa.entel.cl/nuestra-compania

[2] Entel. (2022). Memoria Integrada 2021. Revisado en Abril 2022, del sitio web: https://informacioncorporativa.entel.cl/inversionistas/memoria-anual

[3] Entel. (2019). "La clave para continuar liderando esta industria ha sido esforzarnos por ser fieles al propósito que hemos definido como Entel. Revisado en Diciembre 2022, del sitio web: https://acortar.link/ogdQ3a

[4] Delgadillo, J., Loyola, F. Plaza, M. (2022). Empresa Nacional de Telecomunicaciones - Comunicado reseña anual. Recuperado desde el sitio web de ICR CHile en Mayo del 2022, del sitio web: https://www.icrchile.cl/index.php/corporaciones/empresa-nacional-de-telecomunicaciones-s-a

[5] SUBTEL. (s/f). ¿Qué es SUBTEL?. Revisado en Abril 2022, del sitio web: https://www.subtel.gob.cl/quienes-somos/

[6] Chakraborty, A. (2021). Weight of Evidence (WoE) and Information Value (IV) — how to use it in EDA and Model Building?. Revisado en Septiembre del 2022, del enlace: https://medium.com/mlearning-ai/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-buildin

[7] OEB. (2018). WOE, IV and Scorecards in Credit Risk Modelling. Revisado en Agosto del 2022, del enlace: https://rstudio-pubs-static.s3.amazonaws.com/376828_032c59adbc984b0ab892ce0026370352.html#22_information_value_(iv)

[8] Nykodym, T. et.al. (2018). Generalized Linear Modeling with H2O. Revisado en Septiembre del 2022, del enlace: https://www.h2o.ai/resources/booklet/generalized-linear-modeling-with-h2o/?_ga=2.215109015.1829966795.1666175919-1076031283.1663797839

[9] Amat, J. (2020). Árboles de decisión, random forest, gradient boosting y C5.0. Revisado en Junio 2022, del enlace: https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#C50

[10] Na8. (2017). Random Forest, el poder del Ensamble. Revisado en Junio 2022, del enlace: https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/

[11] Bentéjac, C. et. al. (2021). A comparative analysis of gradient boosting algorithms. Revisado en Junio 2022, del enlace: https://link.springer.com/article/10.1007/s10462-020-09896-5

[12] Amat, J. (2020). Gradient Boosting con Python. Revisado en Junio del 2022, del enlace: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

[13] Amat, J. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Revisado en Junio del 2022, del enlace: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

[14] Hoens, T. R., Chawla, N. V. (2013). Imbalanced Datasets: From Sampling to Classifiers. Imbalanced Learning, 43–59. doi:10.1002/9781118646106.ch3

[15] Mellado, B. (2021). Redes Neuronales para predicción de pago de deudas de clientes de una empresa de retail financiero. Revisado en Junio del 2022, del enlace: https://repositorio.uchile.cl/handle/2250/180310

[16] Cañadas, R. (2021). Curvas ROC. Revisado en Junio del 2022, del enlace: https://abdatum.com/machine-learning/curvas-roc

[17] Library H2O. (sin fecha). Performance and prediccion of classification model. Revisado en Agosto del 2022, del enlace: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/performance-and-prediction.html?highlight=AUC#classification

[18] Burke, M. (2018). Population Stability Index. Revisado en Agosto del 2022, del enlace: https://mwburke.github.io/data%20science/2018/04/29/population-stability-index.html

[19] Aravena, S. (2022). Entel planea para 2022 su mayor inversión en cuatro años: más de US$660 millones. Revisado en Abril del 2022, del sitio web La Tercera: https://www.latercera.com/pulso/noticia/entel-planea-mayor-inversion-desde-2018-con-foco-en-negocios-movil-y-hogar/AX2HNQ7JU5DLLOYUMRTRYFQFLM/

[20] Entel. (2022). Entel planea invertir más de US 660 millones en 2022. Revisado en Abril del 2022, del sitio web:https://informacioncorporativa.entel.cl/comunicados-de-prensa/posts/entel-planea-invertir-mas-de-us-660-millones-en-2022-con-fuerte-foco-en-los-negocios-m

[21] Moine, J., Gordillo, S. & Silvia, A. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. Revisado en Junio del 2022, del enlace: https://host170.sedici.unlp.edu.ar/server/api/core/bitstreams/4df570ab-4869-4d71-ad9f-120094302c50/content

[22] Chapman, P. & et. al. (2007). Step by step data mining guide. Revisado en Junio del 2022, del enlace: https://www.studocu.com/cl/document/pontificia-universidad-catolica-de-chile/business-analytics/crisp-dm-resumen-en-espanol/8584573

[23] A. Peréz J. Segundo. (2018.) El análisis económico y financiero que has de realizar en el inicio de un proyecto. Revisado en Febrero 2022, del enlace: https://emprendedores.uca.es/wp-content/uploads/2018/02/5-analisis-econ%c3%b3mico-financiero.pdf

[24] Niño, M. (2016). CRISP-DM: Fase de "Comprensión del negocio" (Business Understanding). Revisado en Junio del 2022, del siguiente enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-comprension-negocio-business-understanding.html

[25] Niño, M. (2016). CRISP-DM: Fase de "Comprensión de los datos" (Data Understanding). Revisado en Junio del 2022, del enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-comprension-datos-data-understanding.html

[26] Niño, M. (2016). CRISP-DM: Fase de "Preparación de los datos" (Data Preparation). Revisado en Junio del 2022, del enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-preparacion-datos-data-preparation.

[27] Niño, M. (2016). CRISP-DM: Fase de "Modelado" (Modeling). Revisado en Junio del 20222, del enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-modelado-modeling.html

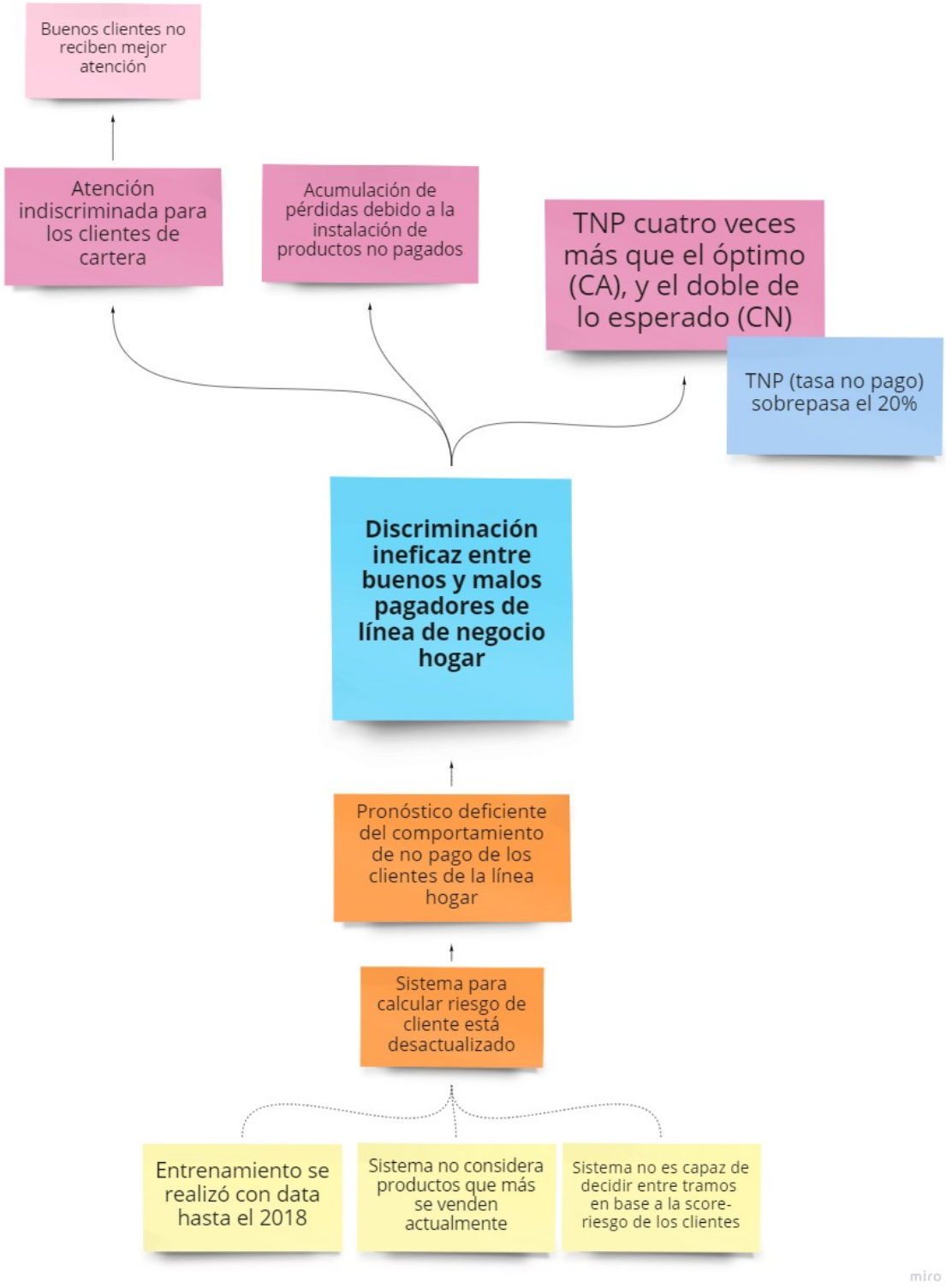
[28] Niño, M. (2016). CRISP-DM: Fase de "Evaluación" (Evaluation). Revisado en Junio del 2022, del enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-evaluacion-evaluation.html

[29] Niño, M. (2016). CRISP-DM: Fase de "Despliegue" (Deployment). Revisado en Junio del 2022, del enlace: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-despliegue-deployment.html

Anexos

Anexo A: Árbol de problemas

Figura 7.1: Árbol de problemas construido. Elaboración propia.



Anexo B: Objetivos y actividades de las diferentes etapas de la metodología CRISP-DM

De acuerdo a una recapitulación de la metodología realizada por Mikael Niño [24], para lograr comprender efectivamente el negocio se deben cumplir los siguientes objetivos y tareas:

- **Determinar los objetivos de negocio:** Comprender al detalle, desde una perspectiva de negocio, qué es lo que el cliente quiere conseguir realmente. El objetivo es descubrir desde el principio factores importantes que pueden influir el resultado del proyecto. Se debe registrar la información que se conoce sobre la situación de negocio de la organización al comienzo del proyecto, así como los criterios de éxito y utilidad del resultado desde el punto de vista del negocio.
- **Evaluar el contexto:** Caracterizar al detalle los recursos (personas, datos, recursos de computación, software, etc.), requerimientos (alcance y calidad de los resultados, así como aspectos de seguridad y legalidad en el uso de los datos), asunciones y otros factores a considerar al determinar los objetivos y plan del proyecto (por ejemplo la gestión de riesgos y planes de contingencia). También se debe realizar un análisis comparativo de los costes del proyecto y los potenciales beneficios para el negocio si el proyecto es exitoso.
- **Determinar los objetivos de minería de datos:** Pasar de la enunciación de los objetivos en términos de negocio a describirlos en el plano técnico, ligado a los conceptos de minería de datos. Al igual que con los objetivos de negocio, hay que determinar unos criterios de éxito técnico e identificar quiénes son los roles dentro del proyecto que van a evaluar el cumplimiento de estos criterios.
- **Generar el plan de proyecto:** Es importante incluir una valoración preliminar del tipo de herramientas y técnicas que pueden requerirse en el trabajo, dado que la selección puede influir en el proyecto completo.

Ahora bien en lo que respecta a la segunda etapa **comprensión de los datos**, Niño indica lo siguiente [25]:

- **Capturar datos iniciales:** Hacerse con los datos (o, primeramente, con la posibilidad de acceder a los mismos) que se han identificado dentro de los recursos clave del proyecto. Se debe realizar una caracterización de los datasets, sus localizaciones, los métodos usados para conseguirlos y los problemas encontrados y su resolución.
- **Describir los datos:** Realizar una caracterización general de los datos obtenidos: su formato, cantidad (número de registros y campos) y cualquier otra característica descubierta en este primer vistazo general. Esta caracterización debe servir para evaluar si los datos obtenidos satisfacen los requerimientos relevantes a este respecto identificados en la fase anterior.
- **Explorar los datos:** Realizar un análisis preliminar de los datos utilizando diferentes herramientas de consulta, visualización y elaboración de informes. En esta exploración nos deberíamos fijar en la distribución de los atributos clave, en las relaciones entre subconjuntos pequeños de los atributos o en las propiedades de determinadas "subpoblaciones" dentro del total de los datos.
- **Verificar la calidad de los datos:** En este examen de la calidad de los datos deberíamos

fijarnos en cuestiones como las siguientes: si están completos los datos (cubren todos los casos que se requieren), si son correctos, cómo de frecuentes son los errores, si hay missing values (cómo se representan, donde y con qué frecuencia ocurren), etc.

Con respecto a la tercera etapa, **preparación de datos**, se tiene que los objetivos y actividades son [26]:

- **Selección de datos:** Decisión sobre los datos a emplear en el análisis, usando criterios relativos a la relevancia para los objetivos, la calidad de los datos o restricciones técnicas. La selección a realizar se refiere tanto a los atributos o campos de los registros del dataset como a los registros en sí.
- **Limpieza de datos:** Se debe "elevar" el nivel de calidad de los datos al requerido por las técnicas de análisis. Esta tarea incluye la inserción de valores por defecto adecuados, o el uso de modelado para estimar los valores ausentes (missing values). Se deben documentar las decisiones y acciones para resolver los problemas de calidad de datos que ya fueron identificados en la fase anterior.
- **Construcción de datos:** A partir de los datos originalmente capturados, se generan atributos derivados, nuevos registros o valores transformados de atributos existentes, en función de los requerimientos para preparar la entrada a las herramientas de modelado.
- **Integración de datos:** Esta tarea se enfoca a la combinación de múltiples tablas o registros para crear nuevos, uniendo por ejemplo datos sobre un mismo objeto pero que se encuentran dispersos en diferentes fuentes, o realizando agregaciones que resumen información contenida en varios registros.
- **Dar formato a datos:** Estas transformaciones se refieren a modificaciones sintácticas que se hacen sobre los datos, sin alterar su significado pero que pueden ser requeridas por la herramienta de modelado a utilizar. Por ejemplo, puede que haya requisitos en el orden de los atributos, o que la herramienta de modelado requiera que los registros estén ordenados según el atributo resultado. En otros casos es necesario presentarlos en un orden más aleatorio del que vienen inicialmente en el dataset (donde suelen tener algún orden determinado).

Para la fase de **modelado**, Niño indica [27]:

- **Selección de la técnica de modelado:** Aunque ya desde el principio del proyecto, en la fase de comprensión del negocio, se realiza una selección preliminar del tipo de técnica a emplear, en este caso la tarea se centra en poner "nombre y apellidos" a la técnica, de entre las diferentes opciones de configuración, versionado, etc. que puede presentar. Además, hay que tener en cuenta que muchas técnicas de modelado funcionan bajo la premisa de unas asunciones específicas sobre los datos (p.ej. distribuciones uniformes, ausencia de missing values, atributos simbólicos para la clase, etc.), por lo que las asunciones realizadas para seleccionar una u otra técnica deben quedar documentadas.
- **Diseño de los test:** Antes de ponernos a generar un modelo, debemos diseñar el procedimiento según el cual se va a medir la calidad y validez del modelo. Esto abarca la métrica concreta de error que se va a emplear, o la descripción del plan para entrenar y evaluar los

modelos, incluyendo el diseño de la separación entre datos de entrenamiento, de testeo y de validación.

- **Construcción del modelo:** Consiste en la ejecución del algoritmo de modelado seleccionado sobre el dataset preparado siguiendo el procedimiento diseñado. Es importante documentar la parametrización utilizada y la justificación de la elección, así como una descripción del modelo resultante, lo interpretable que resulta y las dificultades para dicha interpretación.
- **Evaluación del modelo:** Partiendo de la calidad del modelo o modelos obtenidos según las métricas definidas en el procedimiento diseñado, se realiza también una interpretación y contraste preliminares de los modelos según el conocimiento del dominio y los objetivos de éxito planteados en términos de negocio. La conclusión de esta tarea puede implicar una revisión de la tarea de construcción del modelo para cambiar la configuración de los parámetros de la técnica, y así afinar en la calidad del resultado.

En lo que respecta a la etapa de **evaluación**, las metas y sus actividades respectivas son [28]:

- **Evaluación de los resultados:** Así como los pasos previos ligados a la evaluación se centraban en la precisión y la generalidad del modelo, en este caso la tarea se centraría en medir el grado en el que el modelo cumple los objetivos de negocio y detectar si hay alguna razón ligada al negocio por la que el modelo es deficiente. Se puede plantear también la evaluación del modelo dentro de su aplicación real, si el tiempo y presupuesto lo permiten.
- **Revisión del proceso:** Se debe realizar una revisión más exhaustiva de lo que ha sido el trabajo de minería de datos y los pasos seguidos (si han sido eficaces y eficientes, si admiten mejoras, si podrían haberse planteado con una aproximación diferente), para determinar si hay factores importantes que se han pasado por alto y analizar aspectos de aseguramiento de la calidad de los modelos.
- **Decisión sobre siguientes pasos:** Según las conclusiones de la evaluación de los resultados y de la revisión del proceso, se toma una decisión sobre los siguientes pasos a afrontar: pasar a la fase de despliegue para poner el modelo en operación, hacer nuevas iteraciones de las fases anteriores, iniciar nuevos proyectos de minería de datos, etc.

Finalmente, para la fase de **despliegue**, el mismo autor propone [29]:

- **Planificación del despliegue:** Es necesario determinar una estrategia para la puesta en operación del modelo, identificando los pasos necesarios y cómo (quién, cuándo) ejecutarlos.
- **Planificación de la monitorización y mantenimiento:** Las tareas de monitorización y mantenimiento del modelo puesto en producción son una parte muy importante de la integración del resultado de un proyecto de minería de datos dentro de la operativa diaria de un entorno de negocio. Una buena planificación de estos aspectos ayuda a evitar efectos no deseados, como por ejemplo una utilización incorrecta de los resultados del análisis. Las características del despliegue diseñado en la tarea anterior influyen en cómo debe diseñarse esta tarea.
- **Informe final del proyecto:** Se debe realizar un compendio de los diferentes entregables y documentaciones generadas a lo largo del proyecto, resumiendo y organizando los pasos