UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FISICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERIA INDUSTRIAL

ESTIMACIÓN DE CUSTOMER LIFETIME VALUE EN RETAIL FINANCIERO

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL INDUSTRIAL

MATHIAS ALFREDO WEITHOFER ALBORNOZ

PROFESOR GUÍA:
MARCEL GOIC FIGUEROA

MIEMBROS DE LA COMISIÓN:

ALEJANDRA PUENTE CHANDÍA

PEDRO URZÚA SALINAS

SANTIAGO DE CHILE

2020

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: INGENIERO CIVIL INDUSTRIAL
POR: Mathias Alfredo Weithofer Albornoz
FECHA: 28/09/2020
PROFESOR GUÍA: Marcel Goic Figueroa

ESTIMACIÓN DE CUSTOMER LIFETIME VALUE EN RETAIL FINANCIERO

La memoria se desenvuelve en el sector del retail financiero, en particular en una compañía que posee el 15% de participación de mercado respecto a la cantidad de tarjetas vigentes, representados en más de 1.8 millones de clientes activos.

Se busca encontrar quién es el mejor cliente de la empresa, con el fin de redireccionar las campañas de marketing, aumentar ingresos y mejorar la fidelidad presente en los clientes. Para esto, se ocupará como métrica el Customer Lifetime Value. Esto es, mostrar cuanto valor traerá un cliente a la empresa en el futuro.

El cálculo se hará en una proyección de 12 meses porque se ocupará para la redirección de las campañas que son de formato anual. Se dividió en modelar la actividad del cliente y la proyección de su margen futuro.

Para calcular la actividad de la cartera, se utilizan modelos logísticos dependientes del periodo, donde sólo ira cambiando la variable dependiente. Junto a esto, se le comparará con una metodología Weight of Evidence y un benchmark de un modelo único. Los resultados del modelo logístico muestran que a medida que pasa el tiempo, su accuracy baja de un 84% a un 78%.

Mientras que, en el cálculo de márgenes futuros, se ocupó como método de predicción ARIMA, ARIMAX y el Prophet de Facebook, el cual es un método Generalized Additive Model. Junto a esto, se compara en métricas de MAPE y RMSE con un modelo de panel. Se ocupan en definitiva un ARIMA y el modelo de panel.

Dentro de las principales conclusiones, no se logró el objetivo de satisfacer las dos métricas de valores futuros, pero si el de calcular el CLV de los clientes y su debida identificación, en donde con el modelo de panel fijo se obtuvieron resultados más acertados. Se proponen otros métodos de predicción con el fin de cumplir los objetivos de manera simultánea.

AGRADECIMIENTOS

En medio de una pandemia, es difícil encontrarse en la situación de que no se tiene contacto con el resto de la gente, en particular con cercanos y familia, pero aquí es donde se ha demostrado de toda la bondad de las personas de las que estoy rodeado, y estoy muy feliz y contento de tenerlos a mi lado, aun si no es físicamente.

Primero quiero agradecer a la empresa en la que se realizó la memoria y a todos los integrantes del área de Inteligencia de Clientes por siempre estar disponibles y dispuestos en ayudarme para lograr los objetivos de la mejor manera posible.

En segundo lugar, agradecer a la Universidad y a los profesores de la comisión por la disposición a guiarme en encontrar las herramientas correctas para la realización del tema. En un momento tan crítico para la sociedad como lo es una pandemia, y también en un contexto tan importante como lo es el estallido social, siempre estaré profundamente agradecido de encontrar profesores que están dispuestos a ayudar y escuchar explicaciones.

Luego, agradecerle a mi familia, en especial a mi madre y padre que son los que me instaron a ser partícipe de la Universidad de Chile, guiándome siempre bajo los valores del esfuerzo, la disciplina y la perseverancia para lograr los objetivos que uno se plantea. También, aprendí de ellos a que no hay que perder el foco de lo que se quiere, aun cuando la salud no lo permite. De verdad, muchísimas gracias por ser el pilar fundamental de mi vida.

A lo largo de mi carrera tuve altos y bajos como cualquier estudiante, pero un hecho fundamental que me marcó fue el nacimiento de mi hijo Maximiliano. Criar una vida, entregar valores, trabajar y estudiar al mismo tiempo fue un esfuerzo tremendo, pero no hay nada más satisfactorio que ver los logros que se obtienen al dar lo mejor de uno. Totalmente agradecido de ti y de mi pareja por el regalo que me entregaron, en donde me mostraron que todo el esfuerzo siempre es recompensado y que nunca debo dejar de luchar por ustedes.

También agradecer a mis compañeros de universidad y del colegio por estar siempre para escucharme, y comprobar que a veces las mejores ideas y enseñanzas no son en una sala de clase, sino en el cómo uno se desenvuelve en el día a día. En esta parte, nombrar también a mis compañeros de los equipos de balonmano a nivel universidad y facultad, donde me mostraron que la competencia es más que un trofeo o una victoria, es lograr que un equipo completo logre una sintonía para obtener resultados y, a final de cuentas, la vida laboral es desenvolverse de esta forma.

Mathias Alfredo Weithofer Albornoz.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

INTRODUCCIÓN

Cencosud Scotiabank es una alianza de uno de los mayores bancos del mundo y uno de los retailers más grandes de Sudamérica. Todo comenzó en el año 2003 con el lanzamiento de la tarjeta de crédito "Jumbo Más", creada de la estrategia multimarca de Cencosud, que agrupó a París, Jumbo y Easy. Luego, en el 2011 se inició la implementación a nivel regional de la tarjeta única Cencosud, que permitió el aprovechamiento más efectivo de la marca, y a la vez mejorar las eficiencias operativas. Finalmente, en el año 2014 se hace una alianza con Scotiabank para reunir la experiencia del retail junto al mundo financiero, de una manera más directa y específica, de donde Scotiabank controla el 51% de las acciones de la alianza corporativa.[1]

La empresa tiene alrededor de 1.8 millones de clientes a lo largo de todo el país, y su público objetivo se ve representado por todo chileno mayor a 25 años. Desde este año, se está pensando en el piloteo de la oferta para los extranjeros residentes en Chile, e incluso un programa para niveles internacionales respecto a la tarjeta de crédito. Existen 4 productos en Cencosud Scotiabank: tarjetas de crédito, créditos de consumo, avances en efectivo y super avances.

La tarjeta de crédito se ve definida como un contrato entre la empresa y el cliente, mostrada como un documento de material plástico para efectuar compras sin pagos en efectivo, e incluso, incurrir en pagos futuros respecto al mismo bien. Suelen tener un límite de dinero para consumir servicios y se puede cargar al cliente un porcentaje por el servicio y, en algunos casos fijos, una cuota fija anual.

Las tarjetas de crédito suelen tener un límite de dinero que permite que la persona compre o consuma servicios, cuando se llega al límite establecido, la tarjeta se inhabilita. No obstante, la entidad emisora de la tarjeta de crédito carga al comerciante un porcentaje por este servicio y en algunos casos una cuota fija anual al tenedor.[2]

Los créditos de consumo son préstamos que las instituciones financieras dan para que el cliente pueda hacerse con el bien consumible que necesite. Suelen emplearse también si el acreedor necesita dinero en efectivo disponible al corto plazo. En general tienen mayor tasa de interés debido a la inmediatez con la que son pedidos. [3]

Los avances en efectivo son las operaciones de la tarjeta de crédito que te permiten obtener dinero en efectivo, ya sea a través de un traspaso de límite de crédito o desde la misma cuenta. Al igual que los créditos de consumo, poseen una tasa de interés alta debido a la instantaneidad con el que se pide el dinero, por lo que se recomienda pagar en pocas cuotas para no aumentar la deuda. [4]

Como último producto se tienen los super avances. Estos son un sobregiro extra en la tarjeta de crédito para pedir efectivo de manera adicional, todo esto más allá del límite que se establezca entre ambas partes. Los intereses serán mucho mayores respecto al cupo inicial.

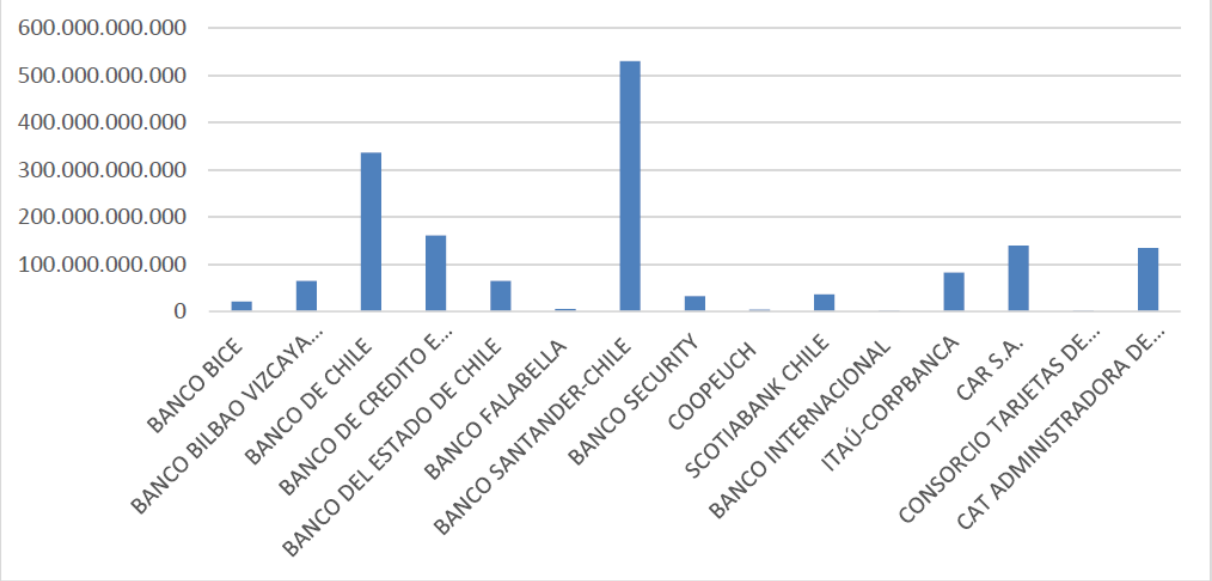
Respecto al dimensionamiento de actividades realizadas por la empresa, se declara que según el último registro del "Informe de Tarjetas de Crédito. Emisores Bancarios" hecho por la Superintendencia de Bancos e Instituciones Financieras Chile (SBIF) en junio de 2018, se ve que, del total de tarjetas de crédito vigentes hasta esa fecha, se tenía el 17,16% del mercado, con un total de 2.209.788 tarjetas (con un universo de 12.876.736). [5]

Dicho esto, se sumarán los 4 productos de Cencosud Scotiabank para ver cuál es el valor que aporta cada cliente a la empresa; con estos valores, que se reportan de manera mensual, se calculará el CLV de estilo truncado para poder orientar los esfuerzos del área de Marketing.

CONTEXTO DE MEMORIA

El mercado financiero, según el "Informe de Tarjetas de Crédito. Emisores Bancarios" citado anteriormente, explica que el mercado financiero de tarjetas de crédito en el mes de julio de 2019 tuvo un total de 33.287.518 transacciones, representadas en un monto de operaciones de $1.607.108.117.570 pesos chilenos. En particular, el retail financiero representado por CAT Administradora de Tarjetas S.A. alrededor de 4 millones de operaciones bancarias, evaluadas en $133.813.303.522 pesos chilenos, representando así un 8.33% del total monetario.

Gráfico 1: Representación de los montos operacionales de los bancos en Chile para julio 2019.



Fuente: Informe de Tarjetas de Crédito. Emisores Bancarios.

Cencosud Scotiabank posee el segundo programa de fidelización más grande de Chile (los puntos Cencosud) el cual se vio reflejado en los más de 2 millones usuarios activos durante el año 2019.

Para contextualizar el tema de memoria, el tiempo de vida del cliente (CLV por sus siglas en inglés) es un término que se usa para determinar el valor que un cliente puede aportar a un negocio durante toda su vida útil dentro de ella. En otras palabras, refiriéndose al tema del alumno, es cuanto podrá aportar un cliente a la empresa durante su tenencia de la tarjeta de crédito.

En general, con esta métrica se busca responder 4 preguntas esenciales: [6]

1. ¿Se está invirtiendo lo suficiente para adquirir a un cliente?
2. ¿Cómo se proyectarán los nuevos clientes de la empresa?
3. ¿Cuánto es lo máximo que se puede gastar para mantener al cliente?
4. ¿Cuál es el valor de la cartera de clientes?

De manera preliminar, todas estas preguntas son necesarias para el proceso que está viviendo la empresa con respecto a la transformación digital y el servicio de excelencia para los clientes, tanto nuevos como antiguos.

Existen distintas categorías y usos del CLV. Para empezar se tienen los que buscan encontrar el impacto de los programas de marketing en la adquisición de clientes, retención o expansión, el segundo tiene que ver con encontrar cual es el factor más importante que influye en el CLV de un cliente (como puede ser la retención, o que a mayor tiempo en la empresa no significa ser más rentable), un tercer tipo ocupa el CLV para relacionarlo con el valor de la empresa, de tal manera que se encuentre una correlación entre el valor de la firma empresarial con lo entregado por sus clientes.

Junto a estos usos, existen distintas metodologías para encontrar los valores del CLV. Se ocuparán modelos de series de tiempo junto a una predicción de actividad futura del cliente, con el fin de saber cuál será la rentabilidad esperada del cliente en una ventana de 12 meses y así calcular su CLV.

JUSTIFICACIÓN DEL PROBLEMA

Se busca como objetivo el reconocer el comportamiento de los clientes, con el fin de predecir sus acciones y tomar acción respecto a los resultados. La empresa vive un proceso de transformación digital, en donde todas las decisiones que ha tomado la gerencia de Inteligencia de Clientes son siempre respecto al comportamiento pasado de los clientes. Es decir, todos los resultados que se le entregan a la gerencia de Marketing y Segmentos van con la base de que la forma en la se desenvolvió el cliente en el pasado será de la misma forma en el futuro.

La gerencia no está cumpliendo la visión que se está implementando en la compañía, ya que las decisiones en base a datos sólo se cumplen respecto al comportamiento pasado, pero no respecto a cómo serán estos clientes en el futuro.

Junto a esto se agrega que, al estar enmarcado dentro de un mercado financiero, la probabilidad de retención de los clientes es baja, en donde menos del 30% de los clientes tiene una recencia menor a 4 meses, por lo que todo lo que han construido respecto al pasado no tiene una claridad de que funcionará con lo que pase en los próximos años.

Dicho esto, la empresa se ha mantenido en un rango determinado de descuentos y acciones en el rubro durante este tiempo sin lograr segmentar las mejores acciones para cada tipo de cliente y esto es porque no saben si esto traerá mejores beneficios en el futuro, o si el cliente se irá aun cuando tenga el beneficio pre-entregado. Ocupando el CLV se tendrá una métrica para continuar las acciones o modificarlas según el resultado de cada cliente.

Entonces el cálculo de valor de cliente se justifica con las distintas decisiones que se pueden tomar a nivel compañía según el valor que se obtiene de cada cliente, de tal manera que se direccione los presupuestos de las campañas marketing según se estime conveniente; de hecho, se espera del área de Inteligencia de Clientes que logre conocer a cabalidad el comportamiento de los clientes fieles, y además lograr predecir el ingreso de nuevos clientes junto a sus características provenientes.

Luego la oportunidad identificada corresponde a calcular cuánto vale un cliente considerando el comportamiento esperado futuro de manera que, a partir de esto, se le pueda ofrecer lo que necesite para que se quede y disminuir la probabilidad de fuga de la empresa. Es decir, se buscará contextualizar mejor a cada cliente con el fin de entregarle mejores beneficios, en especial a aquellos que tengan mejor rentabilidad para la empresa.

Hay que considerar que el poder predictivo de esta métrica no es alto, debido a las diferentes combinaciones de probabilidades y modelos que contienen sus propios errores y márgenes. Pero, lo que le da valor y sentido al cálculo del valor futuro del cliente, es lograr separar bueno y malos clientes respecto a lo que se espera de ellos, y en base a los resultados, tomar decisiones respecto al cómo abordar a cada tipo de cliente y cuanto uno está dispuesto a entregar para captar o mantener a los clientes. [7]

El cálculo de CLV para la empresa significará que podrán redistribuir sus campañas de marketing para los siguientes periodos que sean estimados convenientes. Dicho esto, se buscará responder la pregunta: ¿Quién es el mejor cliente de la empresa? En base a esta pregunta, se desenvolverá el desarrollo de la memoria.

OBJETIVOS

El objetivo general es calcular el valor económico del ciclo de vida de los clientes de un retail financiero, que muestre su comportamiento en un plazo de 12 meses y que apoye la toma de decisiones de la transformación digital.

Objetivos Específicos:

i. Crear una base de datos que contenga toda la información y valores que puedan influenciar en la vida activa de un cliente de la tarjeta de crédito la empresa.
ii. Obtener métricas para el cálculo de su retención y sus márgenes futuros por cada cliente.
iii. Evaluar el uso de las distintas alternativas para el cálculo del valor del cliente y optar por el que se ajuste a los estándares de la empresa.
iv. Validar las metodologías utilizadas para luego proyectar el valor del cliente en un total de 12 meses de vida útil.
v. Analizar los resultados de la metodología y determinar cuáles serán los clientes que entreguen más valor a la compañía.

MARCO CONCEPTUAL

El marco conceptual se dividirá en 2: lo que respecta a CLV como tal y los métodos de predicción a utilizar.

Hoy en día el área de marketing ha mutado y la empresa lo sabe. Antes, la empresa se satisfacía con conocer un poco del cliente para saber cómo llegar a ellos y cuánto es lo que gastan. Hoy en día Cencosud Scotiabank busca una transformación digital y esto conlleva a que ocupará la información del cliente de manera más exhaustiva. De esta manera, reducirá el gasto por cliente para que se ejecute una compra y que sea mucho más efectivo.

Hay que saber diferenciar los cambios estructurales de enfocarse en el cliente versus hacer acciones para sentir mejor al cliente. A modo de ejemplo, grandes empresas como Apple, Costco, entre otras, no se centran en el cliente sino más bien en hacer un buen producto para mantener feliz al cliente. Ser centrado en el cliente es mucho más simple que eso, es tener una orientación fija en su satisfacción y en base a eso, lograr el objetivo de . [8]

Un supuesto imprescindible es que no todos los clientes son iguales y es en esta heterogeneidad donde hay que encontrar el valor de cada cliente. Una vez encontrado el cliente adecuado, se le otorga el servicio que necesite con el fin de fidelizarlo y lograr el objetivo de que el cliente esté contento con el servicio y no con un objeto creado. Esto es lo principal que busca la empresa.

Finalmente, todas las empresas encuentran valor en distintos espacios. Estos pueden ser un margen, tener bajos precios, calidad, etc. Lo importante es encontrar esta estrategia una vez encontrado lo que el cliente espera y quiere del servicio que otorga la empresa.

Para una empresa podría ser deseable no hacerse cargo de algunos clientes o asignar diferentes recursos a diferentes grupos de clientes, y esto no es posible según las medidas financieras agregadas. Entonces como contraparte, el CLV es una métrica desagregada que se puede utilizar para identificar clientes rentables y asignar recursos en consecuencia. Al mismo tiempo, el CLV de los clientes actuales y futuros (también llamado capital del cliente o CE) puede ser un buen indicador del valor general de la empresa.

Existen varios tipos de CLV según el uso que se desee, o bien, según los datos disponibles que se tengan para calcularlo. En el caso de esta memoria, se ocupará el estilo truncado. El CLV truncado es aquel que se calcula con un límite de tiempo específico, se proyecta el valor del cliente con un principio y un fin determinado, por lo que se ignoran los valores residuales. Tampoco se considera que el cliente puede fugarse por completo a lo largo de este ciclo y finalmente se considera una tasa de retención fija a lo largo de los periodos. Primero, están los que buscan encontrar el impacto de los programas de

marketing en la adquisición de clientes, retención o expansión, el segundo tiene que ver con encontrar cual es el factor más importante que influye en el CLV de un cliente como puede ser la retención, o que a mayor tiempo en la empresa no significa ser más rentable, un tercer tipo ocupa el CLV para relacionarlo con el valor de la empresa. [9]

Los formatos para estimar un modelo de CLV son variados. Existen los modelos econométricos que junta los fundamentos del CLV de los modelos probabilísticos, pero agrega estimaciones distintas para las tasas de retención y agrega variables anexas a los clientes. Otra forma son los modelos de persistencia, pero se concentran más en los componentes de adquisición y retención; estos son posibles cuando se tiene presente una gran serie de tiempo de los clientes, así tratarlos como un sistema dinámico, modelos computacionales que se enfocan en lograr los mejores acercamientos en la predicción del comportamiento de los clientes. [10]

En particular, la memoria ocupará una parte de modelos econométricos para enfocarse en la retención y actividad de los clientes en la empresa de retail financiero, pero con la diferencia de que se ocupará un modelo de regresión logístico junto a una aplicación de series de tiempo para calcular los márgenes por periodo de los clientes.

El cálculo de valor de vida los clientes, al ser acotado a un año, se optó por calcular cada parte de la fórmula por separado. Esto significa que se aplicaron métodos para predecir los márgenes futuros de los clientes, para la probabilidad de fuga y la tasa de descuento. En este caso, en vez de ser probabilidad de fuga, es una probabilidad de actividad de clientes y la tasa de descuento estaba calculada previamente por normativas empresariales.

Dicho lo anterior, se hablará de los modelos de predicción para la estimación del CLV. Primero, se consideró a las tarjetas de crédito como un sistema "sin contrato" en un tiempo continuo. Las actividades con contrato son aquellas donde las compras se ejercen con obligación en periodos determinados. Es decir, el cliente debe comprar con seguridad durante los lapsos de tiempo que se fijen en el contrato. En contraste, las actividades sin contrato, como las de la tarjeta de crédito, son consideradas como aquellas donde el cliente fija cuando, donde y como ejerce la compra a la empresa. En este contexto se considera como contractual al tener un producto como seguro o tarjeta de crédito, pero se tratará como uno no contractual debido a que el contrato no impone alguna obligación en el uso de los productos.

De acuerdo con Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi. (2011), el árbol de regresión y decisión son los dos algoritmos más populares utilizados en la investigación debido a su simplicidad efectividad. Por esto,

decidieron probar como se comparaban los resultados con las regresiones logísticas. Después de esto, encontraron que las regresiones logísticas y árboles de decisión realizan el mejor enfoque y dan como resultado que esa empresa logre un nivel relativamente bueno de capacidad predictiva. [11]

Gracias a esto, en la memoria se probarán las regresiones logísticas en los clientes del retail financiero, con modelos basados en la información disponible del cliente y del movimiento de la tarjeta de crédito, debido a que ha funcionado anteriormente en otros trabajos con un parentesco en este tipo de movimiento.

Se ocupa como ejemplo el paper de Glady, Nicolas & Baesens, Bart & Croux, Christopher. (2009). [12] En él, se está bajo el contexto de un banco belga de servicios financieros con una muestra de 10.000 clientes y un periodo de estudio de 9 meses. Se hace importante puesto que explica que lo importante dentro de los métodos de CLV es el cómo se calcula los flujos de capital de los clientes. El cómo calcularlo dependerá de cómo se comportan los datos que se utilizarán y se propone, por primera vez, una función de pérdida sensible a CLV y medida basada en estas clasificaciones del banco belga. Se muestra que los enfoques sensibles al costo logran muy buenos resultados en términos de los definidos medida de beneficio. La diferencia entre este documento y lo que se hará en la memoria, es que aquí se ocuparán series de tiempo para predecir a todos los clientes, mientras que en el estudio citado se ocuparon redes neuronales.
Se concentraron los estudios de series de tiempo en los modelos ARIMA y ARIMAX, con lo que se llega a la conclusión -junto a la empresa- de que un ARIMA puede ser un método alternativo a la hora de proyectar las series de tiempo de los clientes de Cencosud Scotiabank. [13] Al ver que estos métodos han tenido éxito en distintos tipos de series, en la memoria se plantea que estos formatos pueden ser utilizados para predecir los movimientos de los clientes de la empresa.

Por último, se pensó en ocupar un método que fuese más rápido y que pronosticara en otro formato las series de tiempo. Es en base a estos antecedentes que se pensó en el Prophet de Facebook. Es un método de pronóstico de series de tiempo temporales basados en un GAM en la que las tendencias no lineales se logran ajustar a las estacionalidades diarias, semanales y anuales, junto a un agregado de efecto de vacaciones.

Este método funciona con buena precisión en las series de tiempo que tienen fuertes efectos estacionales. Prophet es robusto ante los datos faltantes y los cambios en la tendencia, y generalmente maneja bien los valores atípicos. [14] Por ende, se postula que este método puede ser efectivo para lo que se busca en esta memoria, donde al ser más de 1.8 millones de cuentas, se necesita precisión y rapidez a la hora de predecir las series de tiempo.

METODOLOGÍA

La metodología por utilizar será la Knowledge Discovery in Databases (KDD). KDD Se refiere al proceso de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles mediante una serie de pasos lógicos, con el fin de sacar el máximo provecho a los datos y poder llegar a resultados precisos en poco tiempo.
A continuación, se muestran los pasos a seguir para la metodología [15]:

Etapa 1: Selección.

- Para desarrollar el proceso, primero se deben seleccionar los datos necesarios.
- También se determina la información relevante en la construcción.

En este caso, lo primero será seleccionar los datos que influyen en la frecuencia de los clientes, en la cantidad que gastan y lo que influya en la recencia que tengan de usar la tarjeta de crédito. De manera preliminar, se han seleccionado datos demográficos (comuna, ciudad, género), datos de bienes (si posee automóvil, bienes raíces, entre otros) y datos específicos de la tarjeta de crédito (si tiene tarjeta adicional, deudas, etc).

Junto a esto, se tendrá la cantidad gastada del cliente de manera mensual y su frecuencia a las instalaciones donde se pueda ocupar la tarjeta de crédito empresarial.

Etapa 2: Preprocesamiento.

- Exploración de los datos mediante análisis descriptivos.
- Detección de presencia de errores y falta de información.
- Preparación de los datos para el modelamiento.

En esta etapa todos los datos relevantes y la información recopilada se homogeneiza, de tal manera que todos queden con el mismo formato y descripción. Así será mucho más fácil procesar y analizar para los futuros pasos.

Como bien se mencionó en los pasos anteriores, los datos de consumo de los clientes que se tienen son desde diciembre de 2016 hasta enero 2020. Dentro de los primeros datos existe una gran cantidad de missing values que pueden entorpecer el proceso natural que seguirán los datos para el cálculo del valor futuro del cliente en la empresa.

Se da paso a la normalización de los datos y el tratamiento de los valores faltantes dentro de la base de datos entregada por la empresa. Además, se tratan los casos extremos y que no calcen con la lógica (frecuencias negativas, frecuencias que superen los estándares normales, entre otros).

Etapa 3: Transformación.

- Se convierten los datos en información valiosa, es decir, información que explique de mejor manera de problemática.

La limpieza y transformación de los datos tienen el objetivo de mejorar la calidad de los datos entregados del paso anterior y con esto mejora el resultado de la minería en ellos.

En el paso de transformación es cuando se adecuan los datos de tal manera que puedan ser utilizados para los algoritmos seleccionados.

Por ejemplo, graficar y pronosticar de tal manera que el comportamiento de los clientes a la hora de ir a ocupar la tarjeta de crédito sea según Poisson, que la probabilidad de que se mantenga vivo a lo largo del tiempo siga una distribución geométrica, etc. Aquí es donde se prueban todos los supuestos de los modelos con el fin de que se llegue a alguno de los prototipos esperados.

Etapa 4: Data mining.

- Construcción e implementación de los modelos.

Lo que se hace aquí es el descubrimiento de patrones respecto a los datos.

Una vez validado que los datos de los clientes de la tarjeta de crédito se comportan como las distribuciones planteadas anteriormente, se pasa a pronosticar el Customer Lifetime Value de éstos y se proyecta cual será la vida útil de cada cliente. Aquí es donde se corren los modelos de pronóstico de CLV como lo son los Pareto/NBD, RFM, BB/GG, por mencionar algunos métodos de ejemplo.

Etapa 5: Interpretación y evaluación de los resultados.

- Determinación de la calidad de los resultados obtenidos.

En esta última etapa, se analizan los resultados y se evalúa si tienen sentido respecto a lo esperado. Se espera un formato entendible, por lo que las técnicas de visualización son necesarias, ya que los modelos matemáticos pueden ser difíciles en su interpretación para los usuarios que los ocupen.

Una vez entrenados, parametrizados y entregados los resultados de los modelos utilizados, se da paso a la interpretación del comportamiento de los clientes de la empresa. ¿Cuáles son los clientes más valiosos: los que ocupan muchas veces la tarjeta, los que más gastan o los que no se salen del proceso en largos periodos de tiempo? ¿Son resultados esperados los que se presentan o existen anomalías dentro de lo entregado? ¿Cuánto valor aporta lo hecho con respecto a lo que se tenía antes?

Todas estas preguntas serán respondidas al momento de tener los resultados de los modelos de probabilidades entregados por los algoritmos de pronóstico.

ALCANCES Y RESULTADOS ESPERADOS DE LA MEMORIA

De manera preliminar se sabe que se trabajará con todos los segmentos de clientes de la empresa, ya que se espera una evolución del CLV a nivel de cliente individual, en donde se tenga un resultado por cada uno con el fin de poder saber cuánto valdrá cada cliente en el futuro.

Además, se trabajará con los datos de los años entre 2016y 2019, esto para lograr una proyección del desempeño de cada cliente, a lo largo de 12 meses.

En base a esto, el alcance de la memoria es trabajar con todos los clientes activos de la cartera de Cencosud Scotiabank, siendo un cliente activo aquel que aparezca con movimientos respecto al margen que le genera a la empresa a la fecha de enero de 2020.

Esta decisión fue tomada debido a que existen muchos clientes que se van fugando de la tarjeta a lo largo de los periodos, entonces al querer ocupar este cálculo para evaluar su cartera actual y no la cartera perdida, o bien, medir percepciones de cómo atraer más gente a la tarjeta, el margen es la variable que define si un cliente sigue o no usando los productos de la empresa. Por lo tanto, se dejarán afuera del proceso a todos aquellos que ya dejaron de ser clientes o que no hayan aparecido con registro de margen durante enero 2020 (este valor puede ser positivo, negativo o cero).

Los resultados esperados son encontrar las distribuciones de comportamiento de los clientes de la empresa, cálculo del CLV de cada cliente de la cartera de los que usan la tarjeta de crédito de Cencosud Scotiabank, análisis de los resultados encontrados por el cálculo del valor del cliente junto a su respectiva información dentro de un documento y finalmente una clasificación de los clientes en base a los resultados entregados, de tal manera que se sepa cuáles son los mejores clientes proyectados para la empresa y que se puedan tomar decisiones en base a estos resultados. De esta manera la empresa podrá sacar conclusiones para diseñar de una manera efectiva y con foco en el cliente su estrategia de targeting para sus campañas de marketing, se puedan alinear como área con la nueva visión que tiene la empresa de transformación digital y logren tomar decisiones en base a resultados futuros.

Una de las grandes dificultades que pueden surgir a la hora de desarrollar la propuesta de CLV es que las condiciones comerciales actuales pueden influir en lo que genera cada cliente. Para evitar este tipo de problemas se optó por no realizar un CLV con datos reales de 2020, además se decidió por ocupar el año 2019 como datos de testeo y no los últimos datos que se actualicen del año 2020, pero considerando una precaución con los últimos periodos de ese año debido al estallido social que se vivió en el país.

DESARROLLO METODOLÓGICO

Desde la empresa se pidió que el cálculo fuese único para cada cliente. En base a esto, se descartaron desde un principio todos los métodos que usaran clusters para segmentar los clientes según características similares.

Dicho esto, se dividió el cálculo del CLV en 2 partes: una probabilidad de fuga anexada a cada cliente, la rentabilidad esperada del cliente para los siguientes 12 periodos. Hecho esto, se trae el valor del cliente a valor presente con el uso de la tasa de descuento.

$$CLV_i = \sum_{t=1}^{12} \frac{Rentabilidad_{i,t} * \ ActividadDelCliente_{i,t}}{(1+r)^t}$$

En la primera parte, al querer tener un cálculo por cada cliente, se optó por realizar una serie de tiempo, con el fin de que se obtenga cada periodo predicho y este se multiplique por la probabilidad de retención y, junto a esto, un modelo de panel con efectos fijos.

Para la segunda parte, al existir estacionalidad en los periodos dentro del negocio y, agregando la gran incertidumbre que se vive en el actuar de los clientes desde el estallido social de octubre 2019, junto a que no existe una definición de fuga dentro de la empresa, se decidió por hacer un modelo de regresión logística homogéneo por cada periodo en el que se predice el CLV de los clientes, es decir, existirá un modelo genérico por cada mes durante los 12 meses que se proyectará el valor. Con este modelo, se obtendrá una probabilidad de que un cliente compre con la tarjeta en cada mes de predicción. Se comparará con un modelo único que dependa de los periodos, que servirá como benchmark para comparar los resultados de predicción.

Finalmente, la tercera parte respecto a la tasa de descuento será entregada por la cartera de proyectos de la empresa, por lo que será un valor fijo que no dependerá de cálculos del memorista.

Modelos de predicción de probabilidad de retención

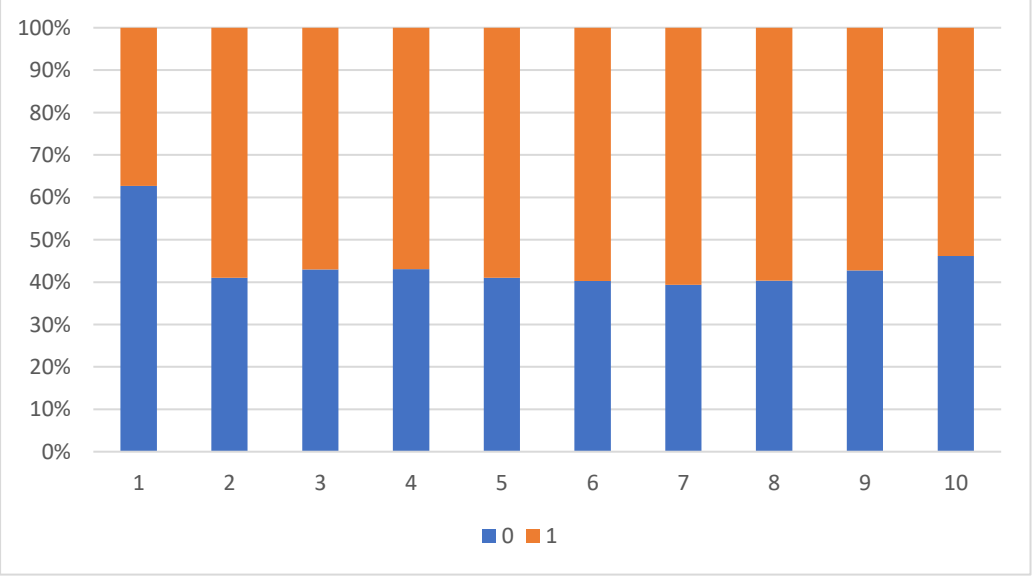
Etapa 1: Selección y estudio de variables.

La base de datos se distribuye en filas donde se encuentran el rut encriptado, junto a sus datos transaccionales con la empresa y sus datos correspondientes al SBIF y demográficos, desde diciembre del año 2016 hasta enero del presente año. Es decir, se tienen alrededor de 2.4 millones de clientes por 38 periodos de estudio (entre diciembre de 2016 y enero de 2020) dando un total aproximado de 90 millones de datos.

Al momento de elegir las variables para modelar, se tenían a disposición 100 variables entre las que se incluían variables de fidelidad, de comportamiento del cliente en la empresa, variables fijas de cada cliente y variables de deudas. Al no tener la capacidad computacional para encontrar las mejores variables a través de un LASSO u otro método similar, se optó por realizar el estudio del comportamiento de las variables candidatos en contraste con la variable dependiente del modelo logístico homogéneo mediante deciles de clientes.

A modo de ejemplo, se muestra la variable línea de crédito disponible para el cliente, perteneciente al grupo de variables de deudas entregada por la Super Intendencia de Bancos e Instituciones Financieras (SBIF).

Gráfico 2. Comportamiento de la variable línea de crédito disponible respecto a la variable dependiente.



En este gráfico se ve el movimiento de la variable dependiente a través de cada percentil de clientes. Si se demostraba un movimiento significativo dentro de los percentiles de la variable, de más de 10% por cada uno, se concluye que la variable podría ser incluida en el modelo homogéneo.

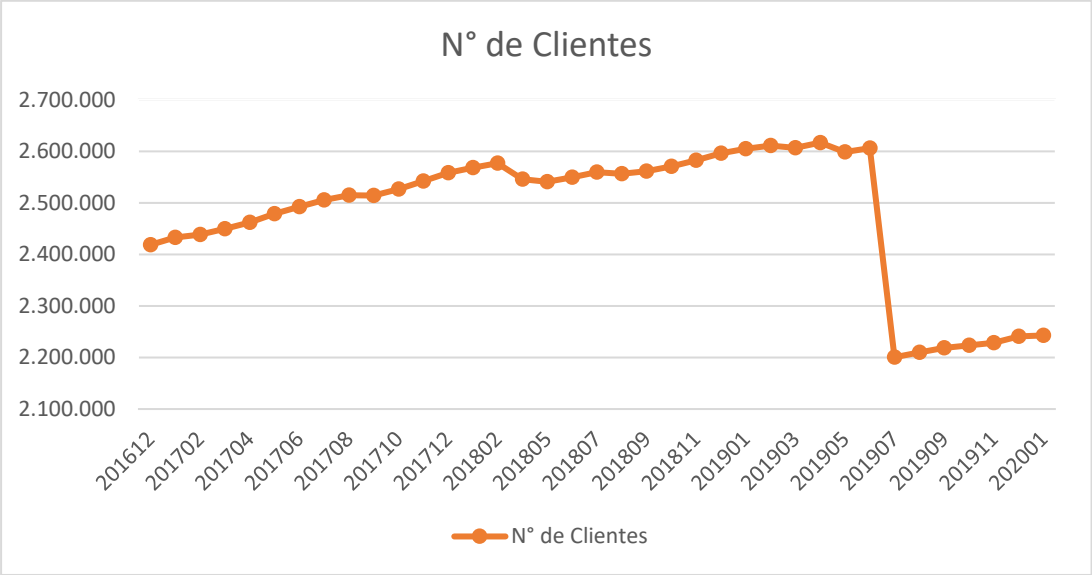
Hecho esto, al ocupar el modelo homogéneo en todos los periodos y tener un menor número de variables -haciéndolo viable computacionalmente- se eligen las variables significativas.

Luego de este estudio, se concluyeron como variables candidatas las siguientes:

1. Ingresos por mes. Estos ingresos se pueden descomponen en margen, compras con tarjeta e ingresos mensuales. La diferencia entre estos es que el primero es la diferencia entre los ingresos y los costos asociados, la segunda solo simboliza las compras que se realizan con la tarjeta de crédito de la empresa y la última representa el ingreso del cliente por periodo. Son variables continuas

   Para trabajar, se ocuparán los datos de ingreso por periodo y de margen. El segundo no se ocupará ya que lo que se busca con este CLV es el valor del cliente en los 4 productos: tarjetas de crédito, créditos de consumo, avances en efectivo y super avances. En promedio, se tienen 2.4 millones de datos de ingreso por periodo.

Gráfico 3: Número de clientes por periodo de estudio en la memoria.



Fuente: Informe de Tarjetas de Crédito. Emisores Bancarios.

En este gráfico se puede ver que en promedio la cantidad de clientes es 2.4 millones de datos. El periodo de baja en los clientes es porque hubo una decisión administrativa respecto a los clientes morosos que produjo que la cantidad bajó considerablemente.

A continuación, se presenta una tabla de correlación entre las variables numéricas con la variable dependiente la cual es el ingreso de cada cliente en cada periodo.

Tabla 1.Tabla de correlaciones respecto al ingreso y su definición

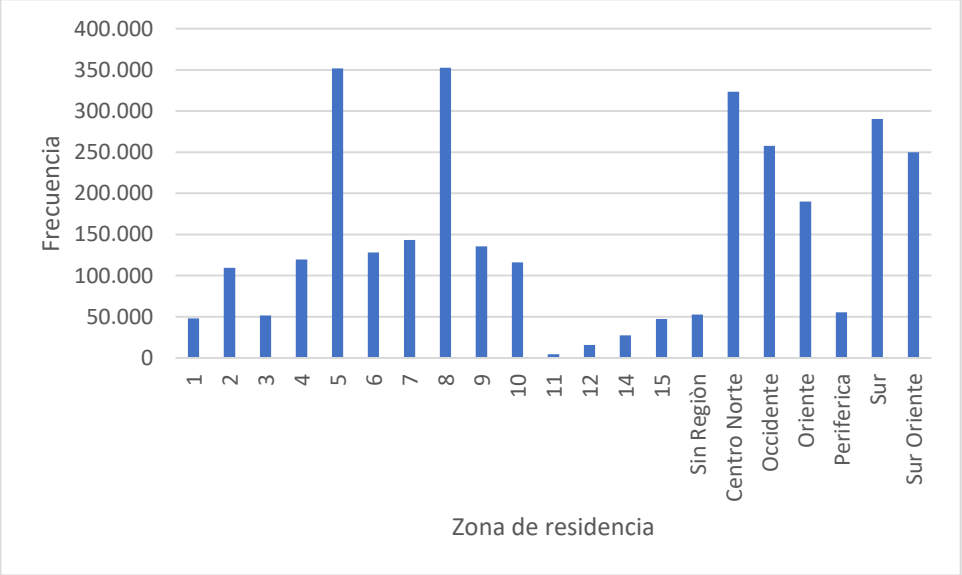
Variable	Definición	Correlaciones
Deuda	Cuanto debe el cliente al retail financiero.	0.53
PorcentajeUsoTarjeta	Porcentaje de uso de la tarjeta de crédito durante el periodo.	0.31
Frecuencia	Veces que ha comprado el cliente.	0.177
SBIF_NumeroInstituciones	Cantidad de instituciones financieras a las que está inscrito el cliente.	0.17
TomaAE	Binaria si el cliente opta por un Avance en Efectivo o no.	0.096
TomaSAE	Binaria si el cliente opta por un Superavance en Efectivo o no.	0.11
SBIF_DeudaVigente	Deuda del cliente a entidades bancarias.	0.05
TomaSegurosTCC	Binaria si el cliente opta por un seguro de tarjeta cerrada o no.	0.04
AntigüedadCliente	Antigüedad mensual del cliente.	0.009
Mes	Mes del año en el que se esté	0.01
SBIF_DeudaHipotecaria	Deuda hipotecaria del cliente.	0.003
SBIF_LineaDeCrédito	Línea de crédito disponible del cliente.	-0.01
TomaSegurosTCA	Binaria si el cliente opta por un seguro de tarjeta abierta o no.	-0.003
Edad	Edad del cliente en el periodo.	-0.005

La descripción de las variables en formato más específico estará en la sección de Anexos. Se continuará describiendo a las variables categóricas.

2. Variables demográficas. Aquí se encuentran las variables del cliente, tales como la región donde vive (en el caso de la región Metropolitana se divide también según la zona donde viva: Centro Norte, Occidente, Oriente, Periférica, Sur o Sur Oriente), el género del cliente. Todas estas variables tienen un retraso de un periodo con respecto al periodo en el que uno se encuentre, es decir, que se está en un periodo t.

2.1. Región-Zona

Gráfico 4: Número de clientes por zona.



Aun cuando se divide la región Metropolitana por sectores, los sectores Centro Norte y Sur tienen casi el mismo peso que la 5ta y 8va región de nuestro país.

2.2. Género

Están incluidos todos los clientes que han comprado en los periodos establecidos, incluyendo aquellos que ya no pertenecen a la cartera de la empresa hoy. Más de 1.6 millones de clientes son representados por el género femenino, siendo así un 55% del total, teniendo así 1,3 millones de clientes de género masculino. Son alrededor de 250 personas que no se identifican con algunos de los dos géneros.

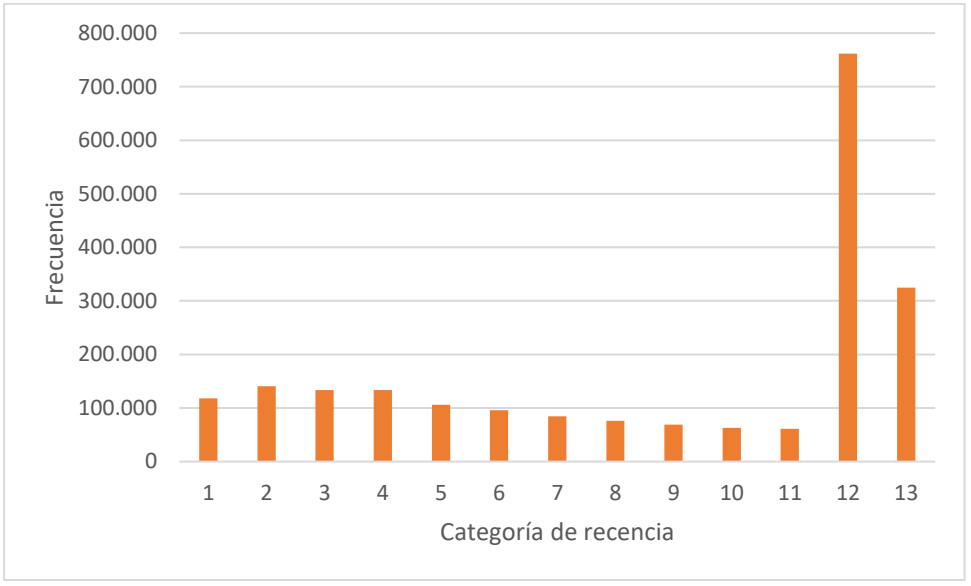
3. Recencia del cliente. existen dos variables anexas: en el retail o con la tarjeta de crédito. Estas variables representan cuanto se demora un cliente en volver a comprar. Una vez que el cliente compra, esta variable vuelve a 0 (variable categórica). Por explicaciones comerciales, esta variable después de 12 meses de recencia se transforma en un valor 99, que significa que la recencia es mayor a la visión que le interesa a la empresa de cuando compran los clientes.

Gráfico 5:Máxima recencia del cliente en el retail.

El número 99, como se ha mencionado, significa que el cliente ha superado el número de meses sin comprar que le interesa al retail financiero, por lo que todos los que superen los 12 meses sin comprar, pasan a esta categoría.

Como se puede ver en el gráfico, el mayor peso de la cartera puede pasar 12 o más meses sin volver a comprar en el retail. Los siguientes puntos altos son los 3 y 4 meses, que son los consideran de manera interna como un corte de fuga de cliente.

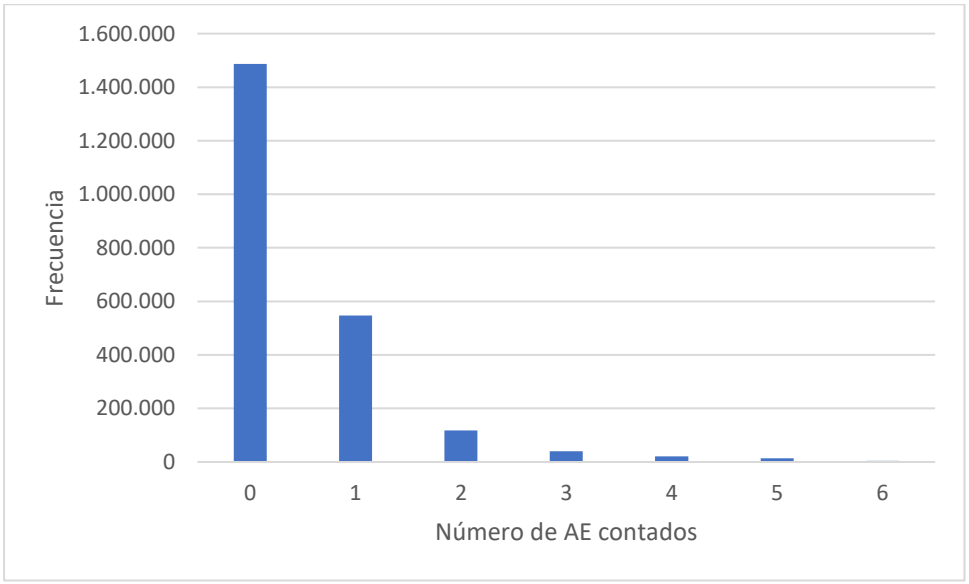
Gráfico 6: Máxima recencia del cliente por tarjeta de la empresa.



Si bien tiene un peso similar los 12 y 99 meses de categoría al igual que en el retail, en este caso se puede ver que existen más clientes que tienen una menor recencia, es decir, que ocupan mucho más la tarjeta en comparación a lo que compran en el retail.

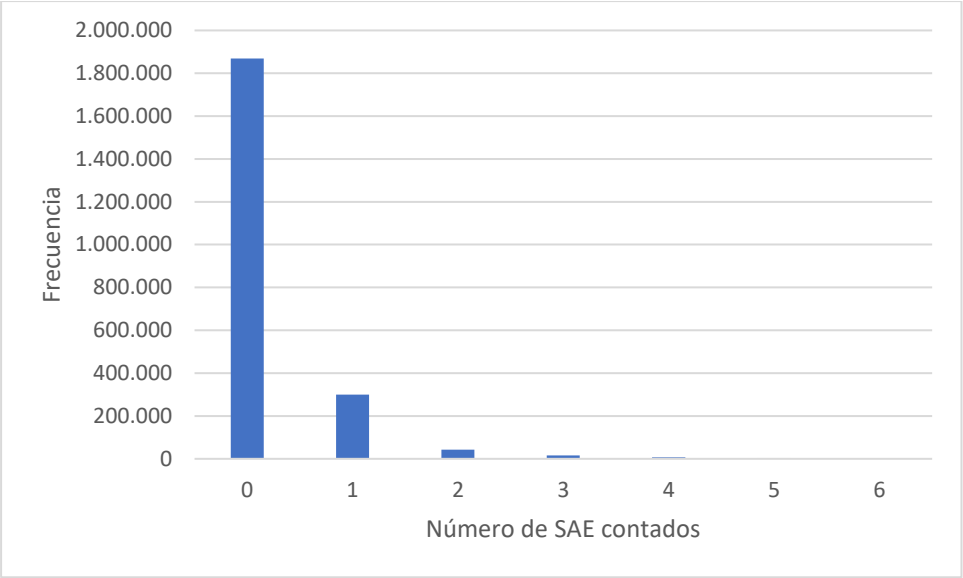
4. Toma de seguros o avances: en esta variable dummy se muestra un 1 cuando el cliente en el mes adquirió un seguro de tarjeta, un avance en efectivo, entre otros productos comerciales. (variable categórica).

Gráfico 7: Toma de avances en efectivo (AE).



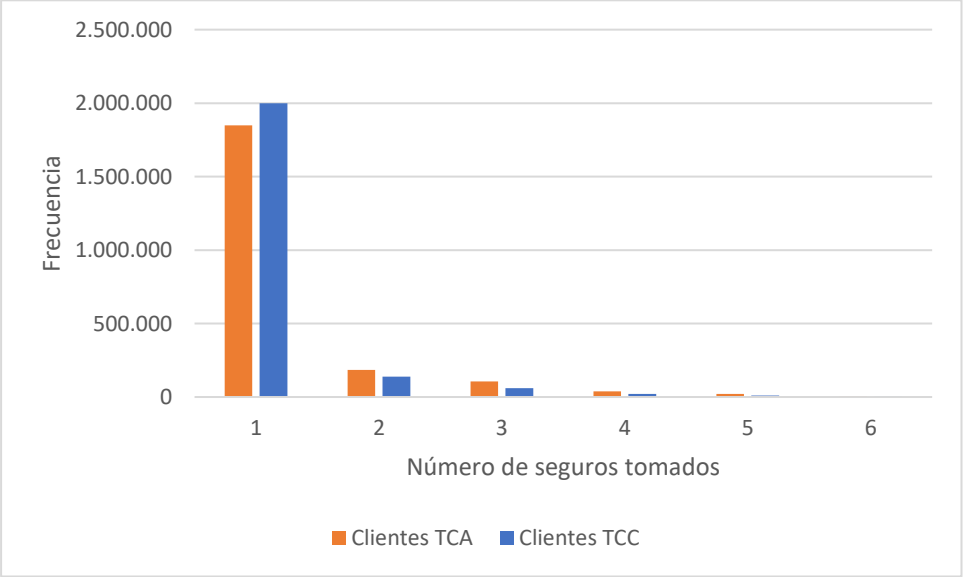
Según el gráfico, más de 1.4 millones de clientes nunca ha tomado un avance en efectivo. Sin embargo, alrededor de 600.000 han comprado entre 1 o 2 productos de esta categoría. Existen clientes que han comprado hasta 50 en el periodo estudiado.

Gráfico 8: Toma de super avances en efectivo (SAE) de los clientes.



La diferencia entre el AE y el SAE es que el segundo es un nuevo préstamo en efectivo sobre el que ya se efectuó. En esta categoría se logra encontrar que más del 75% de los clientes nunca ha tomado un SAE y tan solo 300.000 clientes (12% app) lo ha hecho.

Gráfico 9: Toma de seguros por parte de los clientes de la cartera.

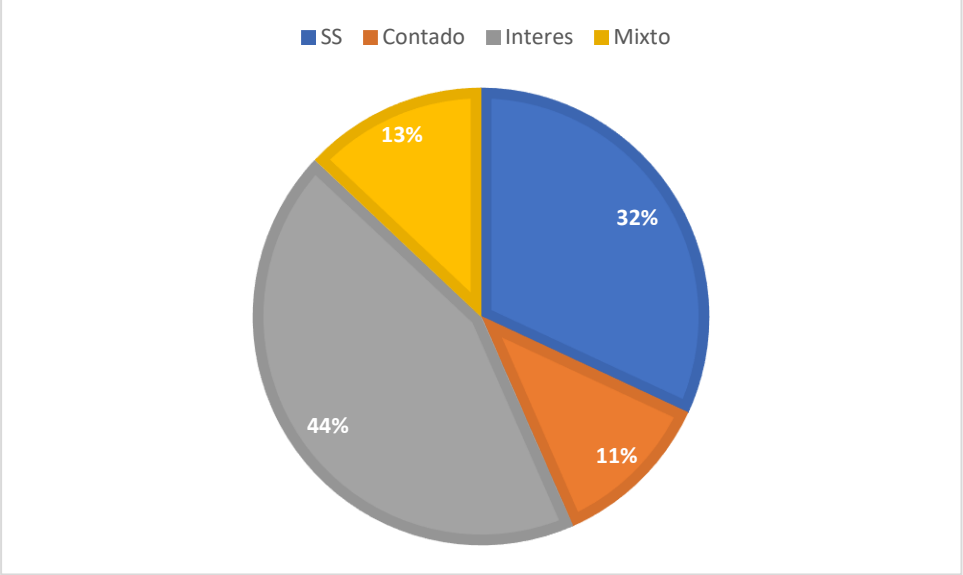


En esta variable categórica se vio el máximo de seguros que ha tomado la cartera de clientes. Más de 1.8 millones en ambas categorías nunca han adquirido uno de estos, y tan solo en promedio 150.000 clientes lo han adquirido una vez.

5. Forma de pago: Se tiene la forma de pago del cliente: contado, crédito o mixto.

   Al igual que en las variables demográficas, estas variables tienen un retraso de un periodo con respecto al periodo actual.

Gráfico 10. Forma de pago más usada por los clientes de la cartera de la empresa.



En el gráfico se puede apreciar que la forma más usada para pagar por parte de los clientes de la cartera de Cencosud Scotiabank es el método de Interés. Cabe destacar que esta variable es difícil de captar, por lo que tiene un porcentaje bastante alto de missing values (representados en el valor SS).

Para tratar este retraso en las variables, se creó una columna nueva en la base de datos que sea de un retraso de uno o dos periodos según corresponda, y se cruzan los ingresos del periodo en el que se encuentre con las variables por tipo de retraso, sea de un periodo (como lo son las variables demográficas) o de dos como lo son las variables SBIF.

Etapa 2: Preprocesamiento y transformación.

Al querer que el análisis de CLV sea por cliente, la empresa pidió explícitamente que la actividad fuese individual, es decir, por cada cliente, entonces es indispensable el cómo se trata los datos que no estén disponibles. El tratamiento de missing values (NA) y de los datos fuera de rango fueron conversados con la empresa y se llegó a un consenso en cada variable:

I. Todo ingreso que no esté se le asigna un 0 puesto que significa que no existió actividad. No existieron NA's en este paso.
II. Toda variable SBIF que sea un NA se le imputará un 0. Eran alrededor de 140.000 datos por periodo que eran de este tipo, representando un 6% por periodo del total de datos. (Entonces, son 140.000 datos por 38 periodos disponibles).
III. En las variables demográficas: en género se crea una categoría nueva llamada "Sin Sexo", en región se crea una categoría llamada "Sin Región" y en edad se le entrega la edad promedio del promedio en el que está el dato faltante.
IV. Al igual que en la edad promedio, respecto a la antigüedad mensual del cliente se le da el promedio del periodo en el que se encuentre; si no existe la toma de seguro se impone un cero, al igual que en el porcentaje de uso de la tarjeta o de la deuda, ya que se asume que no ocupó un seguro, la tarjeta o que no tiene deuda.
V. Para la variable forma de pago se le agrega una categoría llamada "SS" que define que el cliente pago de una forma que no está considerada en los formatos dichos anteriormente, o bien, se desconoce.
VI. Finalmente, para la variable recencia todo NA se agrega a la categoría 99, tanto para recencia retail como para recencia de la tarjeta

Una vez que se trató los missing values, se da paso a la normalización de las variables. Esta normalización fue en dos partes: la primera tuvo que ver con la normalización de las variables continuas y otra de variables categóricas.

Respecto a la primera, se optó por una normalización min-máx. Se trata de que cada dato en la variable continua se le resta el mínimo del dato y se divide por la diferencia del máximo y el mínimo valor. De esta forma, el mínimo será 0 y el máximo un 1 y todos los datos estarán en ese rango.

$$V_{scaled} = \frac{v - \text{mín}\,(v)}{\text{máx}(v) - \text{mín}\,(v)}$$

Para las variables categóricas se procedió a utilizar el método One Hot Encoding (OHE). Este método crea una variable binaria para cada valor de la categoría, con el fin de que se pueda normalizar el peso de cada valor en la categoría y no sea que un valor se esté llevando todo el peso de la información. [16]

Una vez estudiadas las variables, se da paso al estudio de los modelos para el cálculo de probabilidad de retención por periodo en la empresa de retail financiero.

Además, se decide que se ocupará como comparación un modelo de Weight of Evidence (WoE). Lo que hace entregarle un valor fijo a cada valor por categoría, con el fin de que todas las variables categóricas sean transformadas a un valor numérico discretizado. A la vez, para las variables numéricas continuas se hace un paso de discretización, con el fin de agrupar estos valores y poder asignarles el corte correspondiente. De esta forma, se logra correr un modelo de regresión logística, pero con cortes numéricos en las variables.

En resumen, el tratamiento que se le hizo al WoE fue el siguiente:

1. Para una variable continua, se dividen los datos en 10 partes (o menos, dependiendo de la distribución).
2. Se calcula el número de eventos y no eventos en cada grupo (es decir, los bins de cada variable).
3. Se calcula el porcentaje de eventos y de no eventos en cada grupo. En el caso de la memoria, el evento es si hubo un ingreso mayor a 0 en el periodo o no.
4. Se calcula el WOE tomando la división logarítmica natural entre los porcentajes del cálculo anterior, es decir:

$$WoE = \ln \frac{Eventos\ que\ no\ ocurrieron}{Eventos\ que\ ocurrieron}$$

Se optó por comparar el uso de la regresión logística mediante la normalización utilizada versus los resultados del WoE porque en la empresa se trabajó antes con este tipo de modelos y se quiso ver que tanta ganancia entregaba la regresión logística en comparación al modelo de evidencia. Para comparar, se verán los resultados de las matrices de confusión tanto para los datos de entrenamiento como los de testeo.

En resumidas cuentas, se hicieron 2 normalizaciones y luego se compararon los resultados de las regresiones logísticas por ambos caminos y se opta por el que entregue un mejor Accuracy en la matriz de confusión, con tal de encontrar el mejor método.

Etapa 3: Selección de training y test set.

Aquí se da paso a la selección de rangos de tiempo que pertenecerán a los grupos de entrenamiento y testeo del modelo de retención de clientes por periodo para la empresa del retail financiero.

Se llega al consenso con la empresa de ocupar todo el año 2018 como dataset de entrenamiento para los modelos, mientras que el año 2019 y parte del 2020 parte del testeo. La argumentación para elegir estos años es que durante el 2017 el negoció mutó mucho en la forma de actuar respecto a sus clientes, y la toma de decisiones en ese entonces puede sesgar los resultados de los modelos, mientras que el año 2018 y 2019 tuvieron una forma más homogénea respecto a las decisiones empresariales.

Se tendrá un muestreo aleatorio de 1 millón de datos y que este se distribuya en todos los periodos.

Etapa 4: Modelos de probabilidad de retención.
Se decide ocupar una regresión logística en base a lo que se vio durante el estudio del marco conceptual, de tal manera que los modelos serán genéricos para todos los periodos, pero lo que irá cambiando por periodo es la variable dependiente:

$$Y_t = 1\ si\ compra\ en\ el\ periodo\ t, 0\ sino.$$

Donde t= objetivo a t meses proyectados. A modo de ejemplo, la variable dependiente Y_1 corresponde a aquella que intenta predecir si el cliente comprará en el siguiente periodo al que uno se encuentra, Y_7 si el cliente volverá a comprar en el séptimo periodo siguiente, etc.

Dicho esto, el modelo genérico tomaría la forma de la sumatoria de todas las variables, y esto corrió para los t periodos a predecir, es decir, para los 12 periodos:

$$Yt = Variables\ SBIF_t + Variables\ Demográficas_t + Variables\ Cliente_t + \varepsilon_t$$

En donde Variables_SBIF$_t$ = Deuda vigente, deuda hipotecaria, línea de crédito y número de instituciones

Variables Demográficas$_t$ = Edad del periodo, antigüedad del cliente mensual, zona donde vive, mes de compra.

Variables Cliente$_t$ = Porcentaje de uso de tarjeta, deuda con la empresa, formas de pago, recencias.

Las probabilidades predichas de la regresión logit se separaron en el valor 0.5. Si era mayor a este, se le consideraba un 1, sino representaba un 0.

A todos estos modelos, se les realizó una matriz de confusión tanto para el entrenamiento como para los datos de prueba, con el fin de monitorear los errores tipo I y II, junto al cálculo de la curva ROC correspondiente junto a su área bajo la curva. Junto a esto se agregó un gráfico de cómo está separando la variable pronosticada contra las probabilidades entregadas por el modelo, esto se hace con el fin de que el modelo esté separando de manera correcta los valores reales de los datos con respecto a las probabilidades entregadas por el modelo. Estos gráficos detallados de cada modelo se verán en Anexos

A continuación, se mostrarán los resultados de las variables independientes que predicen para el mes siguientes, los siguientes 4 meses y 9 meses, el resto estarán disponibles en Anexos. Hay que recordar que, al ser modelos homogéneos, lo único que va cambiando es la variable dependiente (ingreso en el mes siguiente, cuarto y noveno mes).

Tabla 2. Resultados de modelos de retención.

	Accuracy train	AUC ROC train	Accuracy test	AUC ROC test
Modelo 1er mes	83,8%	86%	84,2%	86,4%
Modelo 4to mes	79,6%	85,5%	80%	85,9%
Modelo 9no mes	78,8%	82,3%	79,2%	82,3%

Hecho todo lo anterior, a modo de ejemplo, se hizo un gráfico de cómo estaba separando las probabilidades predichas por la regresión logística en comparación a los datos reales. Es decir, se quería asegurar de que el modelo estuviese separando bien las probabilidades del modelo contra los datos reales de la variable binaria.

Gráfico 11: Separación del modelo de target 1 mes.

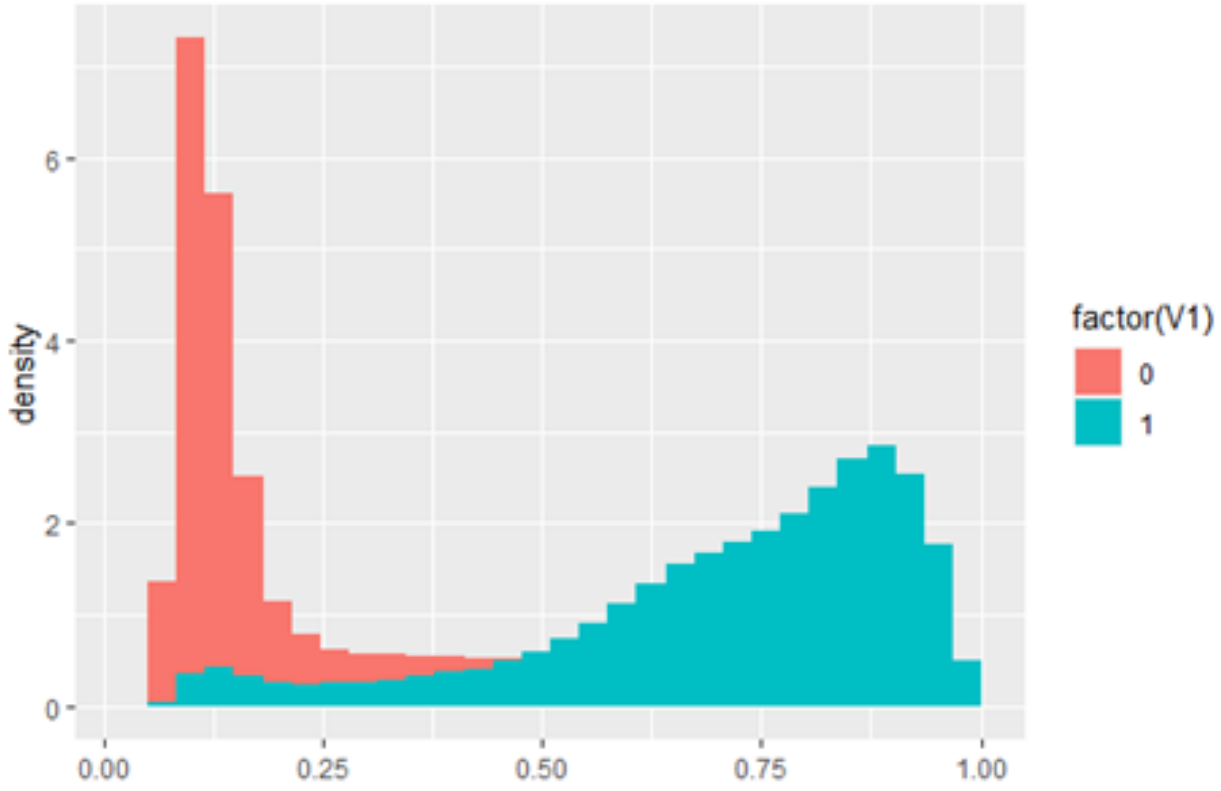
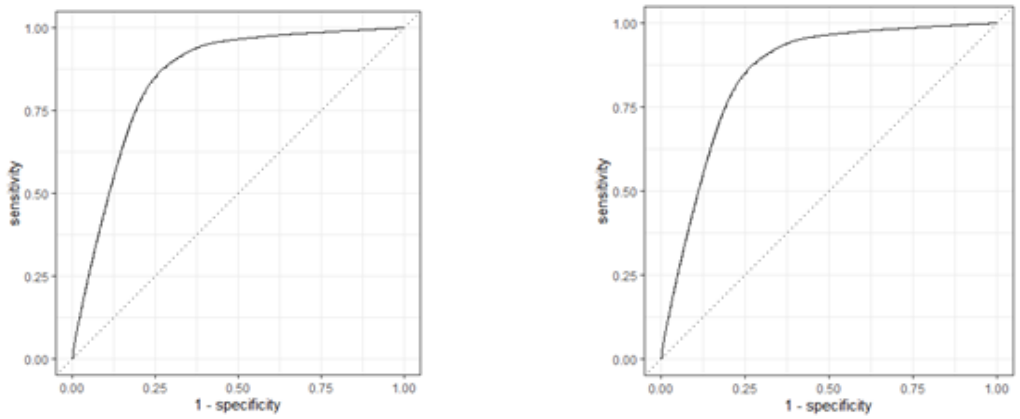


Gráfico 12. Curvas ROC train/test modelo 1 mes.



El área bajo las curvas ROC es de 86% para el train (lado izquierdo) y 85,7% para el test (lado derecho). Claramente, a medida que van avanzandos las variables dependientes este valor va disminuyendo hasta llegar a valores cercanos al 80%.

Con estos resultados, se comprueba que el modelo sabe diferenciar de manera correcta las probabilidades. Se da por concluido el modelo homogéneo para la variable independiente de predicción para el periodo siguiente. Además, en los modelos se repite constantemente que las variables de deuda, antigüedad del cliente y edad son de las más influyentes. En particular, se logra comprobar una de las hipótesis que se tenía donde se pensaba que los clientes en deuda con la empresa tienen una relación positiva con que vuelvan a comprar.

Para comparar este modelo, se ocupó un modelo homogéneo de Weight of Evidence con el fin de ver si aportaba más esta forma de tratamiento a los datos. Cabe destacar que se mantuvieron tanto los datos de entrenamiento como de testeo para todas las comparaciones que vienen en adelante y el tratamiento dicho en la etapa 2 .

De esta forma, no se ocupa el OHE ni tampoco la normalización de las variables continuas, sino que ahora todas las variables pasan por un proceso de cortes (para las variables categóricas), mientras que las numéricas pasan a ser discretizadas para vivir el mismo proceso de corte.

Esta metodología se comparó con 3 modelos distintos. Serán 3 modelos que irán aumentando de a poco el número de regresores. Se tomaron 5.000 series de tiempo, con sus respectivas variables independientes y los cortes para los datos de entrenamiento y testeo fueron los mismos que en primer método, es decir, un OHE y una normalización.

El primer modelo unitario fue una regresión simple en la que la variable dependiente (1 si generaba ingreso en t, 0 sino) dependía solo de la frecuencia de compra del cliente. Se pensó de esta forma para ver que tanto se podría

ganar haciendo un modelo por periodo v/s un modelo que sea unitario y contenga toda la información.

$$Ingreso_t = Frecuencia_t + \varepsilon_t$$

El segundo modelo unitario es un modelo en donde la variable binaria depende sólo del periodo en el que se encuentra y la frecuencia de compra en los 36 periodos y una interacción entre ambos (es decir, 1 si compra en un periodo, 0 sino y se suman).

$$Ingreso_t = Frecuencia_t + Periodo_t + \varepsilon_t$$

El tercer modelo unitario fue el mismo que el anterior, agregándole las variables ocupadas en los modelos OHE/WoE. Se muestran a continuación los resultados del accuracy de los modelos, las matrices de confusión se encuentran en la sección de Anexo.

$$Ingreso_t = Frecuencia_t + Periodo_t + Variables\ SBIF_t + Variables\ Demográficas_t + Variables\ Cliente_t + \varepsilon_t$$

Tabla 3. Resultados de accuracy para modelos benchmark.

Modelos	Accuracy train	Accuracy test
Modelo frec.	71,4%	72,2%
Modelo frec. y t.	71,7%	72,5%
Modelo var. incl.	74,7%	77,9%

La intención del por qué se necesitaba un modelo homogéneo por periodo es que, como se ha mencionado anteriormente, la empresa no tiene una definición o un corte de periodo de cuando un cliente se considera fugado o cuando un cliente pasa a un estado inactivo. En base a esto, se postuló que, a pesar de esto, lo que se necesitaba era una probabilidad de que el cliente no compre en un periodo t+n y eso podía calcularse a través de un modelo homogéneo por todos los periodos, y que este fuese optimizado en cada periodo.

A modo de resumen, se presentan los resultados de los modelos de retención de clientes y su probabilidad de actividad en los periodos siguientes.

Tabla 4. Tabla resumen de entrenamiento de modelos utilizados

Train				
Métricas	Reg. Logística OHE (mean)	Reg. Logística WoE (mean)	Modelo único (básico)	Modelo único (var. Incluidas)
Accuracy	78,1%	78,2%	71,4%	74,7%
AUC	84%	84,5%	77,9%	80,3%

Tabla 5. Tabla resumen de testeo de modelos utilizados

Test				
Métricas	Reg. Logística OHE (mean)	Reg. Logística WoE (mean)	Modelo único (básico)	Modelo único (var. Incluidas)
Accuracy	79,35%	80,25%	72,2%	77,9%
AUC	84,9%	85,9%	75,7%	77,9%

De aquí se puede expresar que el modelo único logra predecir bastante bien el comportamiento de los clientes de la empresa. Sin embargo, los primeros modelos homogéneos de WoE y OHE lograron mejores resultados en casi un 10% por lo que se primó ocupar esta metodología, en particular la de WoE, y a pesar de que sus métricas van disminuyendo a medida que va cambiando la variable dependiente, siguen teniendo mejores resultados que el modelo único.

Además, se agrega que según los resultados entregados no hay presencia de sobreajuste en los modelos debido a que las variantes entre ambos data sets no es significativo (errores que varían entre menos de 1% y 5%).

Modelos de predicción de márgenes futuros

Una vez calculadas la probabilidad de realizar un ingreso de cada cliente en cada periodo, se dio paso al cálculo que generará cada cliente en estos 12 meses de proyección.

De la distinción que se hizo de las variables de ingresos, se optó por utilizar el margen ya que se define como el valor que entrega cada cliente en cada periodo, por lo tanto, esta es la variable que debe estudiarse para calcular los flujos futuros del cliente. Hay que destacar que el valor del margen es una diferencia entre los ingresos y los costos que generan cada cliente.

Tabla 6: Promedio, mínimo y máximo margen por periodo.

Periodo	Margen Promedio	Mínimo valor	Máximo valor
dic-16	$5.539	-$4.931.642	$17.378.986
ene-17	$5.964	-$4.433.238	$10.273.985
feb-17	$7.460	-$4.470.975	$16.024.929
mar-17	$7.683	-$13.409.701	$11.174.794
abr-17	$6.461	-$8.759.106	$10.188.387
may-17	-$4.785	-$7.502.388	$15.825.529
jun-17	$6.651	-$7.888.016	$10.964.106
jul-17	$8.281	-$7.703.837	$9.025.786
ago-17	$6.892	-$9.876.097	$11.755.939
sept-17	$6.763	-$8.467.927	$9.369.333
oct-17	$6.197	-$6.553.882	$12.787.261
nov-17	$6.569	-$13.551.541	$9.924.963
dic-17	$5.310	-$9.720.810	$4.585.389
ene-18	$7.971	-$14.580.187	$9.468.939
feb-18	$6.469	-$9.753.489	$13.511.559
mar-18	$7.410	-$13.523.108	$6.688.221
abr-18	$6.932	-$12.753.906	$5.893.671
may-18	$6.662	-$10.065.954	$5.466.007
jun-18	$6.691	-$52.179.111	$5.398.379
jul-18	$6.781	-$12.235.159	$5.971.735
ago-18	$5.695	-$9.939.412	$4.965.301
sept-18	$6.373	-$13.943.563	$7.318.039
oct-18	$5.771	-$11.950.646	$16.200.000
nov-18	$5.915	-$10.844.404	$5.036.246
dic-18	$3.596	-$30.854.199	$8.100.000
ene-19	$5.639	-$14.109.604	$6.480.000
feb-19	$5.287	-$14.730.641	$6.622.047
mar-19	$5.312	-$12.055.882	$7.250.135

abr-19	$5.563	-$21.106.058	$16.200.000
may-19	$5.807	$19.029.260	$12.669.255
jun-19	$4.832	-$12.647.233	$6.912.594
jul-19	$4.985	-$18.433.571	$9.324.405
ago-19	$5.230	-$16.279.932	$9.563.984
sept-19	$5.082	-$21.095.976	$12.537.133
oct-19	$3.232	-$17.632.654	$6.316.978
nov-19	$4.029	-$19.979.776	$11.942.811
dic-19	$3.081	-$16.225.658	$17.421.472
ene-20	$4.514	-$23.856.707	$10.603.487

Se ocuparon dos tipos de predicciones. La primera consta del uso de series de tiempo para el cálculo de los márgenes futuros para cada cliente y otro que se utilizó un modelo de panel lineal con efectos fijos.

$$Margen_t = Deuda_t + Antigüedad Cliente_t + PorcentajeUsoTarjeta_t + Edad_t + BinPeriodos_t$$

Donde BinPeriodos es una variable binaria para cada mes.

Para comparar su rendimiento se ocuparon dos métricas: MAPE y RMSE. Se utilizó la primera métrica con el fin de ver su poder de predicción respecto a los márgenes de los clientes de la empresa y ver porcentualmente su error, mientras que el RMSE para ver el error medido en capital.

La siguiente tabla muestra el promedio de RMSE y MAPE de las 2 metodologías a utilizar. En las series de tiempo se ocupó más de un algoritmo, por lo que se utilizó el promedio de todos los utilizados para ambas métricas.

Tabla 7: Promedio de métricas en series de tiempo y en el modelo de panel.

Tipo de proyección	Promedio RMSE	Promedio MAPE
Series de tiempo	$162.338	23,98%
Modelo de panel	$53.980	98,23%

En base a los resultados obtenidos, se ve que no se llega a un consenso respecto a las métricas. Al no llegar a un acuerdo respecto a cual se sobreponía a la otra, se darán dos alternativas de cálculo de flujos futuros para los clientes de la cartera de la empresa.

Modelamiento de series de tiempo

Para el cálculo de serie de tiempo por cliente se tomaron en cuenta dos principales aristas: que fuese rápido y que su capacidad de predicción tuviese menos de un 20% de error MAPE. Este valor se consideró como parámetro para ver si un algoritmo era considerado o no.

El análisis del método a utilizar se hizo a través de una muestra de 6.000 clientes que estuviesen distribuidos porcentualmente según los distintos segmentos únicos de la empresa (que son 4: Masivo, Tradicional, Potencial y Preferente, ordenados de menor a mayor valor y de mayor a menor cantidad de clientes). Entonces se calculan las 6.000 series de tiempo y se ve cuál de estos métodos entregas una mejor métrica MAPE y se tomaba en cuenta la rapidez promedio en la que entregaba los resultados.

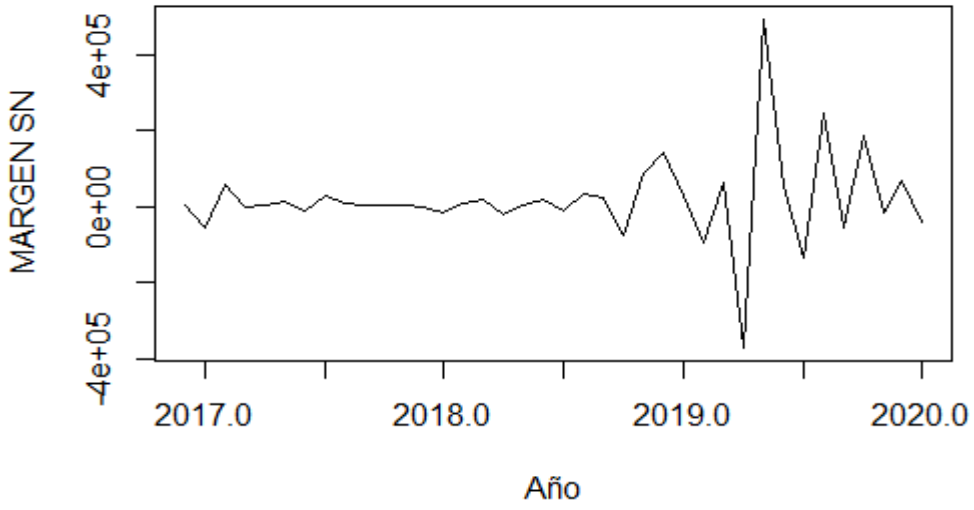
Etapa 1: Normalización de datos de margen por cliente.

Al ocupar una variable como lo es un margen de clientes y mostrado el rango de valores que se tienen, se optó por normalizar los datos de modo que se logre acotar el rango de la variable y logre mejores resultados. Se procedió de la siguiente forma:

$$Margen\_final_{i,t} = \log(\ Margen_{i,t} - \min(Margen_i) + 0{,}0001)$$

Esta normalización se interpreta de la siguiente manera: a cada valor de la serie de tiempo se le restó el mínimo de la serie de tiempo y se le sumó 0,0001, esto con el fin de poder aplicar logaritmo a cada serie sin que tuviese problemas de definición.

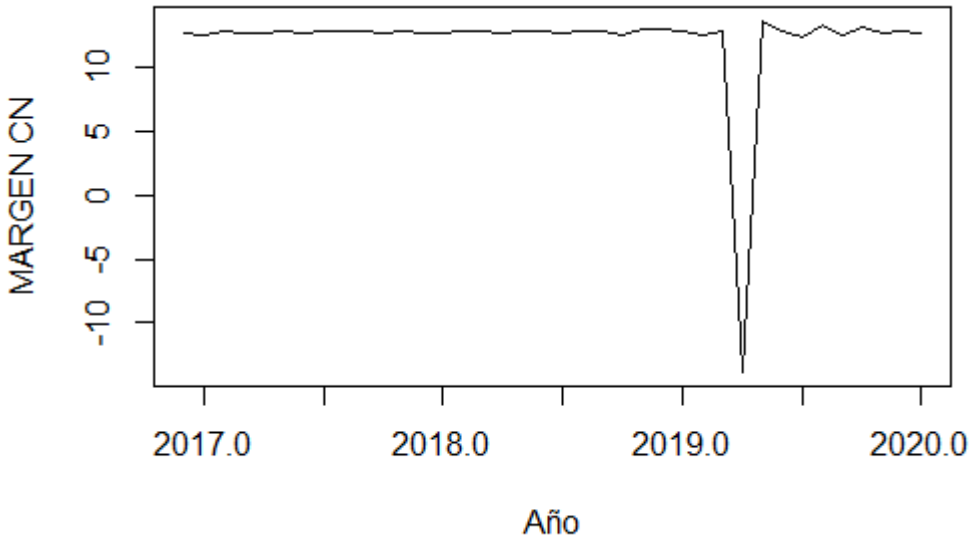
Gráfico 13. Ejemplo de serie de tiempo de cliente de segmento Potencial.



Fuente: Datos obtenidos de base de datos de Cencosud Scotiabank.

Como se ve en el ejemplo de cliente, y como se mencionó anteriormente, los márgenes de los clientes son variables y tienen cotas que pueden afectar al resultado de los modelos. En particular para este cliente del segmento potencial, el rango de margen varía entre -$369.916 y $492.511.

Gráfico 14. Normalización de serie de tiempo de cliente de segmento Potencial.



Fuente: Datos obtenidos de base de datos de Cencosud Scotiabank.

En este caso, se ven los resultados de la normalización de los datos donde ahora los valores fluctúan entre -$13.81 y $13.67, valores mucho más cercanos respecto a los mostrados en el gráfico anterior. Este proceso, como se mencionó anteriormente, se le aplico a toda la cartera de clientes activos de la empresa.

Terminada la sección de normalización, se dio paso a la evaluación de los métodos de predicción de series de tiempo.

Etapa 2: Elección de método de predicción de series de tiempo.
A la hora de buscar el método de predicción, se consideró como factores fundamentales la rapidez con la que procesaba las series de tiempo, la memoria RAM que ocupaba el modelo (debido a que se tenía limites computacionales) y que se adecuara al comportamiento de los clientes. Se ocuparon las series de tiempo hasta diciembre de 2018 como parte de entrenamiento y el año 2019 junto a enero 2020 como test, al igual que en las probabilidades de fuga. De esta forma, se podría calcular la métrica MAPE, que no podía superar un 20% de error promedio en la muestra de 6.000 clientes.

Se ocuparon 3 métodos: ARIMA, ARIMAX y Prophet de Facebook (que es un método GAM).

El primero se consideró debido a la rapidez promedio que tenía por cada serie de tiempo (alrededor de 55 series de tiempo por minuto) y además que cumplía uno de los supuestos de la empresa en la que los márgenes de los clientes estaban correlacionados. Este supuesto se debe a que en general los clientes compran a cuotas o tienen deuda con la empresa, por lo que una compra significa generar margen en n periodos distintos, donde n es el número de cuotas del cliente. Con el paquete de R "Fable", se lograban cumplir todos los tests necesarios de un ARIMA (test de unit-root y encuentra mejor modelo según AIC).

El segundo modelo se puede ver como un modelo de regresión múltiple con uno o más términos autorregresivos (AR) y/o uno o más términos de promedio móvil (MA). Se consideró al cumplir los mismos principios que ARIMA, junto a que tenía un procesamiento aceptable (48 clientes por minuto). Para ocupar ARIMAX, se ocupan distintos regresores que logren explicar de mejor manera la variable margen. Este funciona igual que el paquete anterior, solo que se le suman las variables independientes que se estimen pertinentes.

Se ocuparon, en su mayoría, las variables presentadas anteriormente: porcentaje de uso de la tarjeta de crédito, deuda con el retail financiero y si toma seguros o avances en efectivo. Además de estas, se agregaron variables de deudas SBIF y variables fijas (Antigüedad de cliente y edad) con el fin de abarcar la mayor cantidad de variables.

Finalmente, el Prophet de Facebook fue considerado por ser el más rápido a la hora de predecir series de tiempo (70-75 series de tiempo por minuto). Además, tenía herramientas de calibración de estacionalidad, se le podían agregar factores de vacaciones de la gente, entre otros. El problema de este método es que necesitaba una gran cantidad de periodos para lograr encontrar los parámetros adecuados para cada serie de tiempo.

Tabla 8. Tabla de resultados de la métrica MAPE según método de predicción.

Método	Promedio	Mediana
ARIMA	12,78%	10,35%
ARIMAX*	18,19%	13,67%
Prophet	40,99%	25,09%

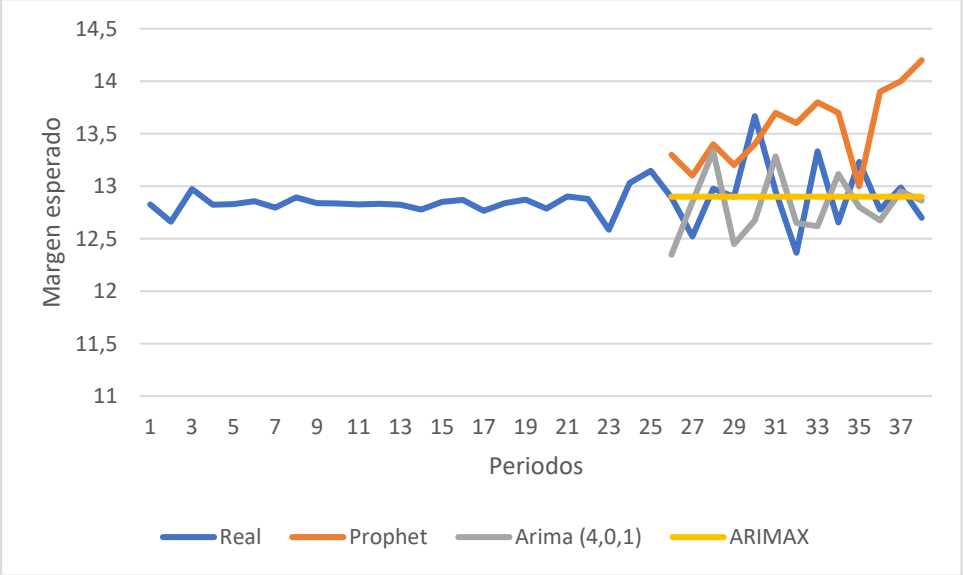
*El resultado del ARIMAX es el promedio y la mediana de todas las combinaciones que se hicieron para este método, el desglose estará en Anexos. Sin embargo, el mejor ARIMAX superó por solo 0,5% a ARIMA y la demora en calcular los resultados fue el doble, por lo que no se consideró viable.

También se les midió según la métrica RMSE, éste representa el error anual por cliente, es decir, que se suman todos los valores dentro de los 12 periodos de testeo y se calcula el promedio y mediana de todos los valores entregados por cada identificador.

Tabla 9. Tabla de métrica RMSE para algoritmos de series de tiempo.

Algoritmo	Promedio RMSE	Mediana RMSE
ARIMA (sin transf).	$300.000	$250.000
ARIMA	$100.900	$45.700
ARIMAX	$98.450	$47.500
Prophet	$150.000	$73.200

Gráfico 15. Ejemplo serie de tiempo con los resultados de cada tipo de método utilizado.



Fuente: Elaboración propia. Datos entregados por Cencosud Scotiabank. En esta serie de tiempo se le imputó el menor valor como un promedio de la serie de tiempo, para ver las diferencias de los métodos.

ARIMAX pide, a su vez, que la variable que se elija como regresora tenga los suficientes niveles (o variaciones) con tal de que se pueda predecir ocupando sus diferencias. Entonces, en ejemplos como este, donde se ocupó como regresor la deuda y el cliente no tenía deuda en específico con la empresa, ARIMAX no logra converger a valores p, d y q. Dicho esto, ocupa un promedio como predicción. Para lograr optimizar este método, tal como se mencionó anteriormente, se debería encontrar la mejor variable por cada serie de tiempo, pero no es factible debido a la gran cantidad de clientes activos.

En base a los resultados anteriores, se optó por seguir con ARIMA como método de predicción de series de tiempo, debido a que obtuvo los mejores resultados y a que su rapidez de predicción es aceptada por la empresa.

Una vez escogido el método para proyectar las series de tiempo por 12 meses, se dio paso a realizar el mismo procedimiento anterior, pero sólo con ARIMA y para toda la cartera de clientes activos de la empresa.

Para esto, se tuvo que cortar la serie de tiempo según antigüedad de actividad en la empresa. Es decir, primero se proyectaron los clientes que poseían información desde diciembre de 2016 hasta enero 2020, después clientes con información desde enero de 2017, y así sucesivamente hasta llegar a los clientes que ingresaron en enero de 2019, después de esto se les considera como "clientes nuevos" y se optó por usar otra forma en ellos.

Para clientes que tuviesen menos de 11 meses de antigüedad, que son reconocidos como clientes nuevos, no se les podía aplicar la misma metodología puesto que no tenían el número suficiente de periodos para predecir la estacionalidad anual que se tenía contemplaba, ni tampoco cumplía los requisitos comerciales de la empresa, por lo que se les aplicó un algoritmo de clusterización Kmeans, en donde se utiliza el CLV de clientes que fueron nuevos un año atrás y, en base a las variables que se tienen de clientes que se ocupan para tratar clientes nuevos, se le asignó el clúster de CLV. Se eligió para trabajar debido a su rapidez y simpleza a la hora de presentar resultados.

El algoritmo K-means, es el más conocido y utilizado por su simplicidad y ser eficaz. Es un procedimiento de clasificación de objetios en un número de K clusters, en el que K es determinado a priori. Representa la media de sus puntos, y entonces, cada clúster se caracteriza por su centroide. Para elegir el número óptimo de clúster, se tomaron en cuenta 2 formas de calcularlos, la media de ancho de silueta (silhouette) y la suma total de cuadrados (wss). En general, siempre llegaron al mismo valor, pero si se demostraba una diferencia entre ambos, entonces se ocupaba un valor intermedio entre ellos.

A modo de ejemplo, si un cliente ingresa en noviembre de 2019, se le hizo un proceso de clusterización de CLV con respecto a un cliente que haya ingresado en noviembre de 2018, asignándole el valor según variables de la SuperIntendencia de Bancos e Instituciones Financieras (SBIF) y otras de fidelización (cuantas veces ha comprado en Cencosud Scotiabank, entre otros).

Modelamiento de regresión con datos de panel y efectos fijos

Tal como se mostró anteriormente, se consideró el modelamiento a través de una regresión que incluya efectos fijos. Si bien se muestra en la tabla 9 que el promedio de error es de alrededor de $54.000 anuales, la mediana del error es de $5.360 por lo que se consideró apropiado a la hora de predecir a los clientes (es decir, en promedio se equivoca $446,66 pesos chilenos mensuales por cada cliente, lo cual se considera aceptable debido a la complejidad de utilizar la variable).

Primero se hace un estudio de las potenciales variables que pueden influir en el margen de un cliente a través del tiempo, tal como se hizo en los pasos anteriores. Se presenta una tabla de correlación respecto a la variable margen, la matriz completa estará presente en Anexos.

Tabla 10. Tabla de correlación de variables con el margen

Deuda	0.0930
FrecuenciaCompra	0.0812
SBIF_NumeroInstituciones	0.0333
PorcentajeUsoTarjeta	0.0651
AntigüedadCliente	0.0135
SBIF_DeudaVigente	0.0102
Edad	0.0086
SBIF_DeudaHipotecaria	-0.0006
SBIF_LineaDeCrédito	-0.0045
Periodo	-0.0061
BinariaDiciembre	-0.0067

De las variables utilizadas, las que presentan mayor correlación positiva son el porcentaje de uso de la tarjeta, la deuda con la empresa y su frecuencia de compra, todas variables pertenecientes al comportamiento del cliente dentro de la empresa.

Se repite el tratamiento de missing values hecho para calcular la probabilidad de retención de cada cliente, junto a que se agrega una variable llamada frecuencia que cuenta cuantas veces el cliente generó un margen mayor a 0.

También se mantienen los periodos de entrenamiento y de testeo, por lo que se utiliza el año 2018 para entrenar un modelo que logre predecir los márgenes de los clientes en el 2019.

Tabla 11. Regresión de Panel con datos fijos

	Variable Dependiente:
	margen_mensual
MesAntiguedadCliente	−87.668
	(69.197)
EdadPeriodo	−131.478
	(745.125)
DeudaCS	−0.0001
	(0.0004)
PorcentajeUsoTarjeta	5,157.324***
	(1,299.579)
Febrero	−525.470
	(991.789)
Marzo	28.368
	(987.520)
Abril	−110.264
	(983.667)
Mayo	1,318.988
	(980.936)
Junio	−501.696
	(979.257)
Agosto	1,483.550
	(979.417)
Septiembre	−1,010.117
	(980.969)
Octubre	586.083
	(983.537)
Noviembre	439.933
	(987.170)
Diciembre	−763.230
	(991.843)
Observations	249,931
R^2	0.0002
Adjusted R^2	−0.042
F Statistic	3.011*** (df = 15; 239916)
Note:	*p<0.1; **p<0.05; ***p<0.01

Para predecir los periodos siguientes (es decir, desde febrero de 2020 hasta febrero de 2021) no se tenía disponible alguna función que lograra proyectar hacia periodos siguientes. Por lo tanto, se hizo una transformación de este modelo de panel con efectos fijos a una regresión logística a través de una

degradación de los datos respecto al promedio (OLS estimation with demeaned data*)* con el fin de lograr construir las proyecciones de los clientes para el 2020. [17]

Una vez que se obtuvieron las proyecciones, se hace un paso atrás en donde se le agrega el promedio degradado a cada proyección del cliente y se obtienen los márgenes futuros.

Al igual que en la metodología anterior, se hizo con una transformación logarítmica intentando mejorar los resultados, sin embargo, estos no lo fueron respecto al hacerlo sin tratamiento, pero si fueron mucho más estables en su RMSE. A pesar de esto, se concluye el uso de este método con la variable sin transformación.

Tabla 12: Resultados de modelamiento de panel respecto a la métrica RMSE.

	Promedio de error RMSE	Mediana de error RMSE
Con transformación	$21.520,04	$19.941,73
Sin transformación	$53.980	$5.360.

Tasa de descuento

Como se mencionó anteriormente, la tasa de descuento no es un cálculo que depende del memorista, sino que fue un cálculo realizado por la empresa.

Gracias al WACC de la compañía, se calcula una tasa respecto los resultados que se esperan del 2020.

$$Ko = Kp * \frac{P}{V} + Kb * (1 - t_c) * (\frac{B}{V})$$

Con un costo de capital (Kp), o tasa de retorno esperada por los accionistas, de un 21%, un costo ponderado de 2,7% (Kb), un capital de 208.357 (P) y una deuda de 1.194.003 (B), un total de activos de 1.402.360 (V) y una tasa de impuestos de un 27% (Tc).

En resumen, la tasa de costo relevante fue de un 5,34% anual, convirtiéndolo en un 0,43% mensual para uso del CLV.

Finalizadas estas 3 etapas, lo que se hace es aplicar la fórmula de Customer Lifetime Value: es agrupar identificador de cliente, multiplicar cada margen por probabilidad de fuga del periodo, y dividirlo según el periodo que sea en su tasa de descuento para calcular el CLV.

RESULTADOS Y CONCLUSIONES

El entregable final es una tabla que contiene los identificadores de cada cliente con su respectivo CLV. Con las series de tiempo se logró pronosticar esta métrica de toda la cartera, mientras que con el modelamiento con efectos fijos se hizo con 200.000 clientes que representa alrededor del 13% de la cartera. Se toma una muestra de 100.000 clientes de las series de tiempo y de la regresión de efectos fijos y se comparan resultados.

Tabla 13. Promedio y mediana de CLV según metodología utilizada.

	Promedio CLV	Mediana CLV
Series de tiempo	-$257.928	-$44.264
Regresión con EF	$53.986	$5.361

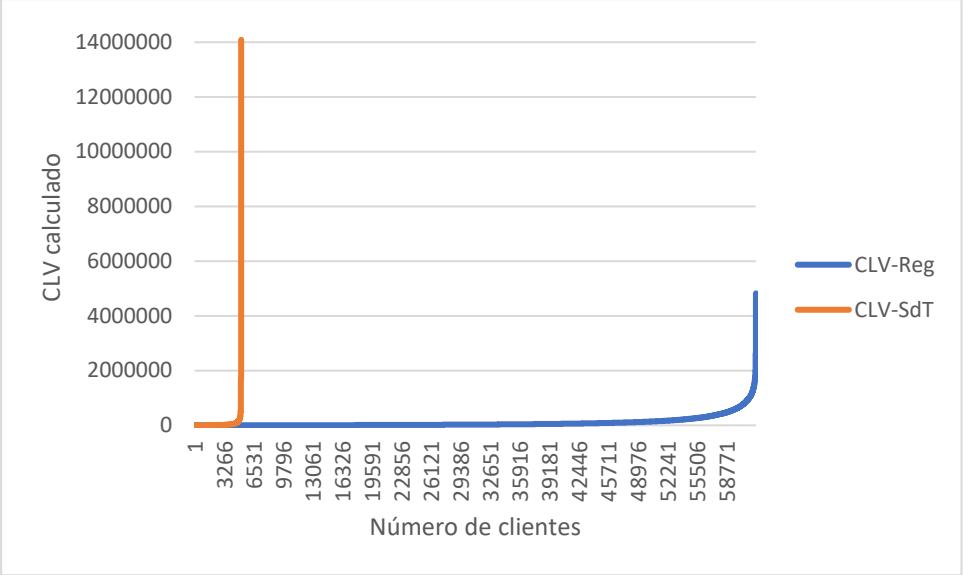
Respecto a los resultados de CLV sólo se enfocará en los valores positivos ya que son los que interesan para el estudio y sus diferencias y, se obtiene que, del total de muestreo, sólo un 5,25% presenta valores positivos en el CLV para la metodología de series de tiempo, mientras que para el modelamiento de regresión es positivo un 62,5%. Se separará el gráfico en CLV mayor a 0 y menores a 0, debido al alto rango que existen de resultados (varían entre -$8.200.000 y $15.000.000).

Gráfico 16. Cota inferior de metodologías de cálculo de CLV.

Aquí se muestran como existen pocos valores de CLV negativo para la metodología de regresión y que se distribuyen en su mayoría en valores menores a -$2.000.000, mientras que para las series de tiempo es una

distribución mucho más esparcida y con crecimiento mucho más lento, donde no se logra distinguir una densidad significativa alrededor de algún valor.

Gráfico 17. Cota superior de metodologías de cálculo de CLV.



Se destaca que en valores de CLV mayores a 0, en las series de tiempo se obtiene un promedio de $63.480 y mediana de $11.135 pesos, mientras que con la regresión se obtiene un promedio de $103.271 y una mediana de $27.850. Se hará un análisis con respecto a los clientes que cumplen esta condición.

En las series de tiempo se obtienen menos resultados positivos y también se encuentran las cotas superiores de CLV. A pesar de estas cotas, de este segmento el 95% de los clientes poseen un CLV menor a $ 2.000.000,con una cota superior de alrededor de $14.000.000. Mientras que en el uso de la regresión con efectos fijos se vuelve a presentar una distribución con un crecimiento más lento, pero no llega a valores tan altos como la otra metodología. El 97% de los clientes posee un CLV menor a $2.000.000 y su cota superior es de $4.823.000.

Con respecto al uso del método de ARIMA, se encontraron muchos clientes que entregaron como resultado ARIMA (0,0,0). Este resultado se ve explicado a que este tipo de clientes, llamados anteriormente flotantes, son aquellos que utilizan la tarjeta de crédito un par de veces debido a alguna oferta y después no vuelven a ocuparla hasta otro tipo de oferta. Debido a esto, el método no encontró mejor alternativa que entregar un modelo predictivo de un promedio de la serie de tiempo. Lamentablemente este grupo de clientes es el que aumentó la métrica RMSE, debido a que, al traer los resultados a valores económicos, el mínimo margen en general superaba los valores entregados por la serie de tiempo, por lo tanto, si un cliente predecía un exponencial

promedio que fuese menor al mínimo margen, toda la serie de tiempo resultaba negativa.

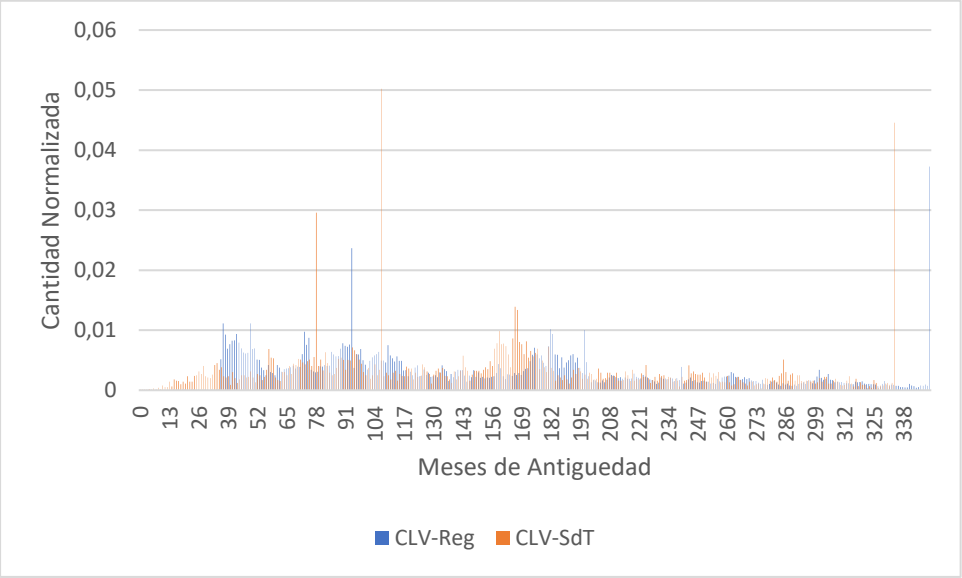
Sin embargo, también están los clientes que están en constante uso de la tarjeta. Lo que más se mostró en este grupo, fueron modelos que no dependían de ninguna diferenciación para ser estacionaria, que ayudó a la rapidez de la proyección de las series de tiempo. También, explica que, si bien los clientes se apegan al comprar un avance en efectivo o un seguro como producto, no quiere decir que el margen que genere ese cliente será influenciable en los periodos posteriores, ya que el margen de generar esa venta estará altamente influenciado, a su vez, por los costos asociados de la acción. Aun cuando la justificación empresarial de que el 5% de la cartera de clientes proyecte un CLV positivo es que está considerando los efectos de la pandemia presente durante el año, sigue siendo exagerado.

Con respecto al modelo de panel con efectos fijos, se encuentra un resultado más realista de lo que pasa con los clientes de la empresa respecto a la metodología de series de tiempo. Tal como se muestra anteriormente, el error promedio anual por cliente en las proyecciones de margen es de alrededor de $2.500 pesos, por lo que el CLV más realista es el que se hizo con esta metodología.

Respecto a las variables que se involucran con el CLV, el género de los clientes, considerando las 2 metodologías, se obtuvo que alrededor del 60% de los que generan un CLV positivo son de género femenino (63,7% en series de tiempo, 58,46% en la regresión) y el restante son de género masculino.

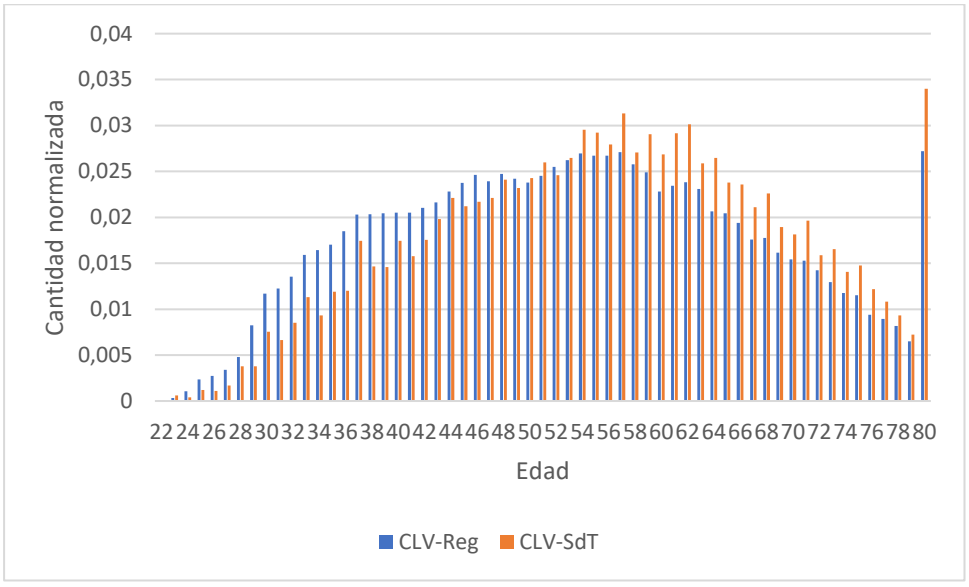
Para la variable de antigüedad de clientes, es posible ver que la concentración de clientes para las series de tiempo es entre los 84 (7 años) y 200 meses (16,7 años) de antigüedad comprando con los productos del a empresa. Mientras que para la regresión de series de tiempo los valores se distribuyen de manera relativamente equitativa a lo largo de la cantidad de meses presentes.

Gráfico 18: Gráfico normalizado por total de clientes con CLV positivo de variable Meses de Antigüedad.



Luego, respecto a la variable edad en el periodo de predicción, se presentaron los siguientes resultados:

Gráfico 19: Edad de clientes con CLV positivo y datos normalizados



Como se puede apreciar en los gráficos, los clientes se distribuyen de manera relativamente similar, sin embargo, existen más clientes que superan los 50 años en las proyecciones de series de tiempo; en los resultados por las regresiones, existe una mayor cantidad de clientes que están entre los 30 y 50 años respecto a la primera metodología, sin embargo, no superan los que son mayores a este rango.

El resto de las variables estudiadas estarán presentes en la sección de Anexos.

Tabla 14. Comparativa de zona entre metodologías CLV

Región o zona	CLV-Reg	CLV-SdT
1	1.45%	1.47%
2	3.56%	2.84%
3	1.68%	1.55%
4	3.89%	3.67%
5	11.80%	11.24%
6	4.00%	3.41%
7	4.58%	4.03%
8	12.29%	13.23%
9	4.51%	4.80%
10	3.53%	3.77%
11	0.14%	0.06%
12	0.44%	0.38%
14	0.73%	0.54%
15	1.62%	1.45%
Centro Norte	9.63%	10.04%
Occidente	8.44%	10.14%
Oriente	6.75%	4.32%
Periférica	1.64%	1.50%
Sin Región	1.29%	1.47%
Sur	9.29%	11.28%
Sur Oriente	8.74%	8.81%

Las primeras conclusiones que se encuentran son respecto al comportamiento de las variables. Es fácil notar que a medida que va avanzando el estudio de las variables es que son muy pocos los clientes que se concentran en el uso de los servicios de la empresa, es decir, son pocos los que usan los seguros, los que tienen una recencia baja. Pero es en estos grupos donde se concentra el comportamiento de que los clientes compren en el periodo, aumenten su porcentaje de uso en la tarjeta de crédito, su deuda con la empresa, entre otros. Es decir, si bien se tiene una gran cantidad de clientes que se podrían llamar "flotantes", también se encuentran los clientes fieles al uso de los servicios de la empresa que muestran un mayor dinamismo.

Además, junto a esto se considera que hubo un tratamiento extenso respecto a missing values presentes en la base de datos. Las principales variables que poseían una abultada cantidad de NA's eran las que corresponden a la recencia

de los clientes, zona donde viven los clientes y métodos de pago. Dicho esto, se puede estar presentando un sesgo respecto a los valores de las variables versus los que sean en la realidad.

Sumado al punto anterior, se explicitó desde un principio que podrían existir sesgos a la hora de proyectar las series de tiempo, donde elegir un punto de inicio para todos podría ser difícil debido a que no todos los clientes poseen el mismo punto de inicio (esto es debido a que no todos los clientes tienen la tarjeta desde el mismo punto, se van sumando y restando clientes desde todos los periodos de estudio). Para tratar este problema, se tomaron todos los clientes que tuvieran algún margen/ingreso en el último periodo de estudio (enero de 2020), de tal forma que se consideren todos los clientes que se mantienen activos dentro de la empresa. Si el cliente presentaba vacíos dentro de las series de tiempo, se rellenaba con ceros, pero de esta forma no se seleccionaban clientes que ya no estuviesen presentes en el uso de los productos de la empresa.

Respecto a los resultados, se logra concluir que el modelo homogéneo logra predecir para todas las variables independientes que se le ha entregado con buenos resultados, variando desde los resultados de la matriz de confusión del primer target de un 84% hasta los últimos períodos que se van acercando a un 77%, el cual sigue siendo un rango aceptable tanto para la empresa como para el memorista.

Cada modelo, eso sí, tiene su forma de cómo lo afectan las variables que se le entregan. Claro es el ejemplo de las variables SBIF, de las recencias o de la toma de avances en efectivo/seguros de tarjeta, que trabajan de maneras distintas a medida que van pasando los modelos, pero también hay variables que se mantienen con su efecto positivo como lo son las variables de deuda con la empresa, el porcentaje de uso y el género masculino, que en su caso tiene un efecto negativo.

Además, si bien el modelo único lograba resultados aceptables de la retención de los clientes, la empresa optó por tener una mejor predicción para los primeros periodos pero que fuese un poco más largo en el proceso, en vez de tener algo más rápido y sencillo como lo que era esta alternativa.

Queda entendido que este método por el cual se está calculando el CLV se hizo debido a que los periodos con los que se pueden trabajar son escasos, pero quedan oportunidades de mejoras incluso en esta forma de trabajo, que son la optimización del modelo homogéneo por periodo y así lograr una mejor precisión a la hora de predecir el comportamiento de los clientes de Cencosud Scotiabank, además de la posibilidad de realizar un modelo unitario.

Para la segunda parte respecto a la toma de decisiones respecto a que método era mejor utilizar, queda claro que es perfectible. Debido a factores externos y la poca capacidad computacional con la que se contaba al principio, no se pudo hacer a través de un método de selección de variable más potente, como lo es en el caso del ARIMAX, sin embargo, el trade-off de los errores del método seleccionado con el tiempo de ejecución logró cumplir las expectativas de la empresa.

Respecto al Prophet, faltaron más periodos de tiempo para lograr calibrar de buena manera el uso del modelo aditivo GAM. En los 38 periodos no fue capaz de encontrar los pesos necesarios para cada serie de tiempo. Además, al ser periodos mensuales no son equitativos a lo largo del tiempo (no todos los meses tienen la misma cantidad de días) por lo que para el algoritmo de Facebook le resulta difícil calibrar un peso ajustado a cada periodo. En base a estos resultados, queda demostrado que no fue posible utilizar esta metodología para pronosticar el margen de los clientes del retail financiero.

A modo de resumen respecto a las series de tiempo y el modelo de panel, no se logró cumplir un enfoque de dos métricas relativamente estables, se logró cumplir con los objetivos de la empresa con respecto a los errores que se encontraran en las series de tiempo y el RMSE respecto a la regresión con efecto fijo. Junto a esto, se obtuvieron probabilidades de fuga en cada periodo por cliente, que ayudará, además de al cálculo del CLV, a plantear nuevas métricas a la empresa, como reconocer nuevos estados de clientes y nuevas segmentaciones. Teniendo un corte que defina cuando cliente es fugado o no, permite ocupar otros métodos de predicción de fuga que no sea por cada periodo que quizás puedan encontrarse con más sentido

Clientes que compraron solo 1 vez una cifra muy alta eran considerados como parte del mejor segmento (Preferente), pero existían clientes que no volvían a comprar en muchos periodos, por lo que plantear un estado de inactivo ayudaría a segmentar de una mejor forma. Esto es, clientes que tienen frecuencias constantes, los que compran poco, pero en grandes cantidades, etc. Tener una segmentación apropiada de clientes, junto al respectivo CLV, puede ayudar a tomar las mejores decisiones de marketing respecto a qué tipo de cliente es, es decir, lograr direccionar las campañas según el tipo de cliente y su comportamiento financiero.

En base a esto, se concluye que se encuentra la métrica de valor futuro de un cliente para Cencosud, mediante los cuales les podrá dar una idea respecto a la toma de decisiones de cuál es la máxima disposición a pagar por un cliente, y a que proyección se espera que se recuperen esas inversiones, donde la más realista es la que se refleja en el modelo de efectos fijos.

ASPECTOS POR MEJORAR

La metodología utilizada tiene aspectos a mejorar significativos:

- No se lograron optimizar los modelos de retención de clientes, por lo tanto, existen variables que no aportan al modelo, pero por falta de tiempo no se logró realizar la minimización del AIC del (los) modelo(s).

- Se planteó como metodología las series de tiempo debido a que se buscaban resultados a nivel de cliente, ya que se quiere encontrar el valor de cada uno dentro de la cartera de la empresa. Se plantearon los algoritmos de Prophet y ARIMA debido a que se tenía de hipótesis que los comportamientos pasados del cliente (no muy lejanos) influían a la hora de generar valor en el siguiente periodo. Por ende, se plantearon algoritmos en donde los valores pasados tuviesen un peso significativo a la hora de proyectar. Sin embargo, se podría hacer una comparación con otro tipo de series, o bien, con unos que sean del mismo estilo pero que sus hiperparámetros sean calculados de forma distinta como lo es un método de Holt, por dar un ejemplo.

- Las series de tiempo son perfectibles. Si bien se lograron buenos resultados con ARIMA, ARIMAX en muchos ejemplos lograba predecir de mejor manera. Sin embargo, al no tener el tiempo para encontrar las variables exógenas precisas para cada cliente, era una metodología inviable. Prophet necesita más periodos de estudios para lograr encontrar estacionalidades, efectos de vacaciones, entre otros influyentes en un año normal. Al tener tan solo 36 periodos de estudios (de los cuales casi 1/3 es para testear) se hizo bastante difícil su uso, junto a que esto lo acompañan los malos resultados.

- Las series de tiempo para el cálculo de CLV permite predecir para máximo 12-18 meses. Por lo tanto, si se quiere esta métrica para un periodo mayor no es viable de utilizar y habría que buscar otras alternativas.

- Para la regresión de panel, se ocuparon como paquete disponible las mismas existentes para hacer el modelo de retención de clientes. Por lo tanto, se dejaron fuera de estudio muchas variables que podrían tener alguna relación con el margen de los clientes de la empresa, se recomienda para un futuro ver si existen otras variables que pueden influir en el uso.

- Éste es el primer enfoque de CLV que da la empresa, en base a esto, esto puede ser abordado desde otras perspectivas más simples (modelo de Recencia-Frecuencia-Monetario) o más complejas como lo es respecto al cálculo real del valor de la empresa. Por lo tanto, otras formas de calcularlo y comparar podrán entregar una mirada con mayor perspectiva respecto al cuáles serán sus mejores clientes.

BIBLIOGRAFÍA

[1] Cencosud. Retail Financiero. Chile. [en línea] <https://www.cencosud.com/cencosud/site/edic/base/port/retail.html> [Consultado entre el 08 y 09 de abril de 2020].

[2] Tarjeta de Crédito" Susana Gil. Economipedia.com. 2017 [en línea] <https://economipedia.com/definiciones/tarjeta-de-credito.html> [Consultado entre el 08 y 09 de abril de 2020].

[3] "¿Qué son los créditos de consumo?" Economiteca.com. 2018. [en línea] <http://economiteca.com/que-son-los-creditos-de-consumo/> [Consultado entre el 08 y 09 de abril de 2020].

[4] "¿Qué es un avance en efectivo?". Julietase. Febrero de 2019. Rankia.cl. <https://www.rankia.cl/blog/mejores-tarjetas-credito-debito/4165546-que-avance-efectivo> [Consultado entre el 08 y 09 de abril de 2020].

[5] SuperIntendencia de Bancos e Instituciones Financieras Chile. 2018. Informe de Tarjetas de Crédito. Emisores Bancarios. [en línea] <https://www.sbif.cl/sbifweb/servlet/InfoFinanciera?indice=4.1&idCategoria=564&tipocont=568> [Consultado entre el 08 y 09 de abril de 2020].

[6] Insight drive organization. Putting data-driven insights to work everywhere, everyday. 2018. Deloitte.com. [en línea]. <https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/insight-driven-organization.html> [Consultado el 20 de abril de 2020].

[7] y [8] Peter Fader (2012). Customer centricity: Focus on the Right Customers for Strategic Advantage. [Consultado entre el 05 y 06 de abril de 2020]

[9] Peter Fader, Bruce G. S. Hardie (2014). What's Wrong With This CLV Formula? [Consultado entre el 21 y 23 de julio de 2020]

[10] (Gupta, Sunil & Hanssens, Dominique & Hardie, Bruce & Kahn, Wiliam & Kumar, V. & Lin, Nathaniel & Ravishanker, Nalini & Sriram, S. (2006). Modeling Customer Lifetime Value. Journal of Service Research. [Consultado entre el 05 y 06 de abril de 2020]

[11] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi. Credit card churn forecasting by logistic regression and decisión tree. [Consultado entre el 05 y 06 de abril de 2020]

[12] Glady, Nicolas & Baesens, Bart & Croux, Christophe. (2009). Modeling Churn Using Customer Lifetime Value. European Journal of Operational Research. 197. 402-411. [Consultado entre el 05 y 06 de abril de 2020]

[13] (Hyndman, R.J., & Athanasopoulos, G. (2019) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3) [Consultado entre el 17 y 18 de abril de 2020]

[14] (Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2) [Consultado entre el 17 y 18 de abril de 2020]

[15] What is One Hot Encoding? Why And When do you have to use it? Octubre, 2019.<https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008fl> [Consultado el 25 de abril de 2020]

[16] Cristina García, Irene Gómez. Algoritmos de aprendizaje: KNN y KMeans. <http://www.it.uc3m.es/~jvillena/irc/practicas/08-09/06.pdf> [Consultado el 10 de julio de 2020]

[17] Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer. (2018) Introduction to Econometrics with R. <https://bookdown.org/machar1991/ITER/10-3-fixed-effects-regression.html> [Consultado el 18 de julio de 2020]

ANEXOS

1. Descriptivo de variables numéricas.

Variables SBIF.

Desde la Superintendencia de Bancos e Instituciones Financieras, institución estatal chilena, se entregan datos monetarios de los clientes respecto a las entidades financieras como lo son: la deuda vigente, número de instituciones financieras en las que está presente el cliente, deuda hipotecaria y línea de crédito disponible. Todas representan variables continuas.

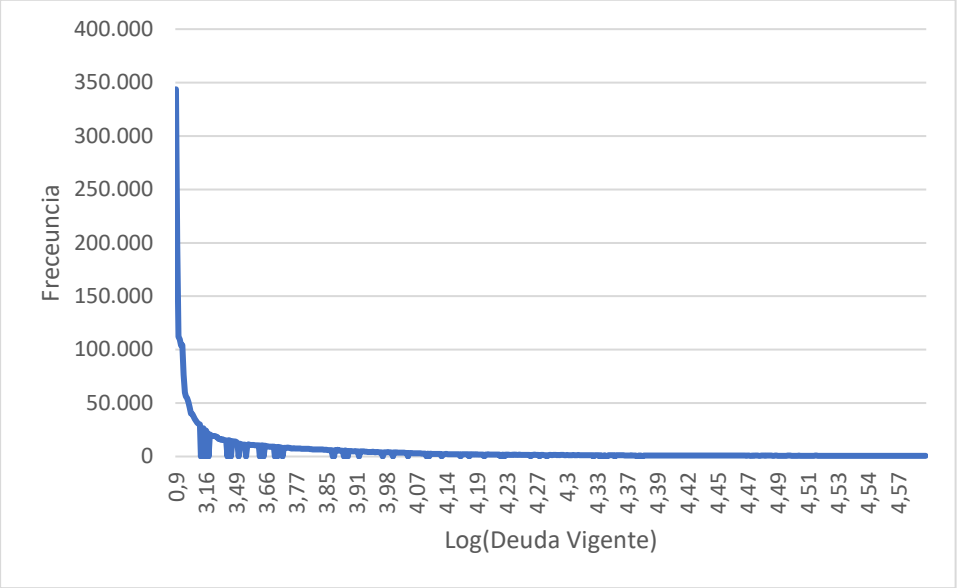
Estas variables tienen un retraso de 2 meses con respecto al periodo actual, por lo que al formar la base de datos se tiene que para el periodo t se tienen los datos actualizados del periodo t-2.

El rango de las variables continuas SBIF entre 0 y más de mil millones de pesos, se optó por hacer un muestreo y se le aplicó un logaritmo para normalizar los valores.

Variable SBIF Deuda vigente (DDIRVG).

Representa la deuda vigente que poseen los clientes por créditos comerciales u otros específicos.

Gráfico 20: Logaritmo de las deudas vigentes de los clientes.



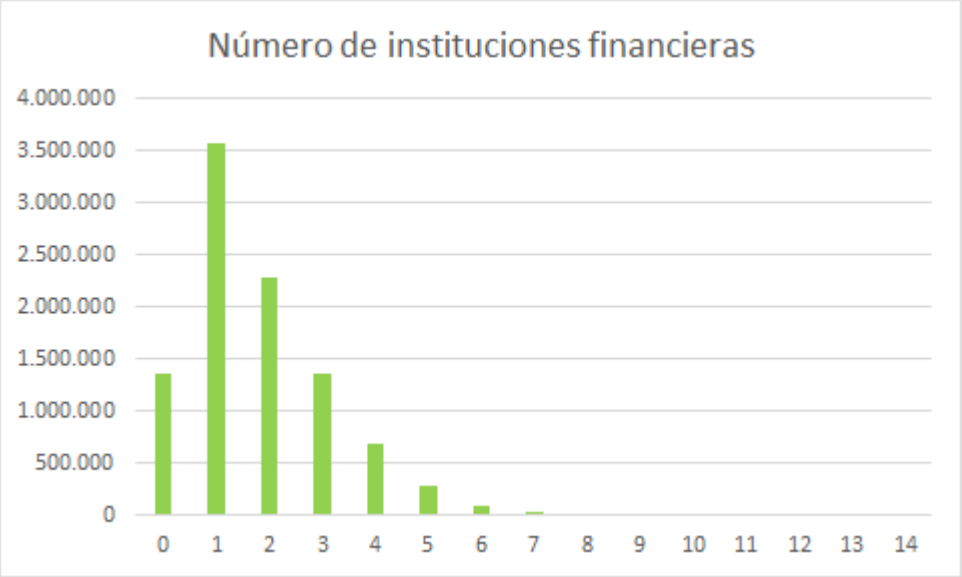
Fuente: Elaboración propia.

Tal como muestra el gráfico, su concentración es en valores bajos de deuda vigente según la SBIF.

Variable SBIF número de instituciones (NINSDD).

Tal como dice la variable, representa el número de instituciones financieras a las que está inscrita el cliente.

Gráfico 21: : Número de instituciones financieras por clientes.



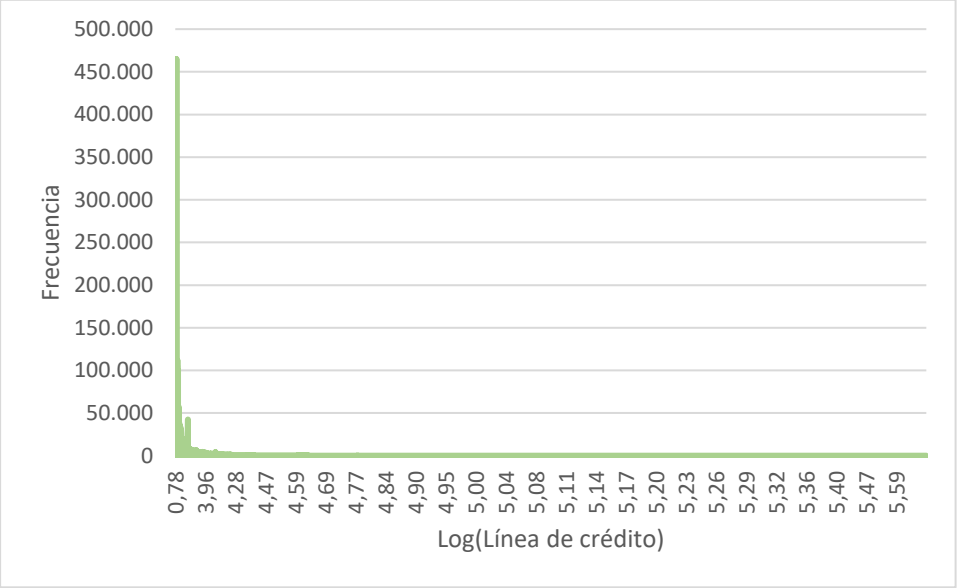
Fuente: Elaboración propia.

Este gráfico representa el máximo de número de instituciones financieras a las que están anexados los clientes.

Variable SBIF deuda hipotecaria (DDAHIP).

La variable DDAHIP es la deuda hipotecaria del cliente para el periodo que está siendo estudiado, considerando eso sí, el retraso mencionado anteriormente. También se le aplicó el logaritmo, para efectos de visualización.

Gráfico 22: Deuda hipotecaria de los clientes.



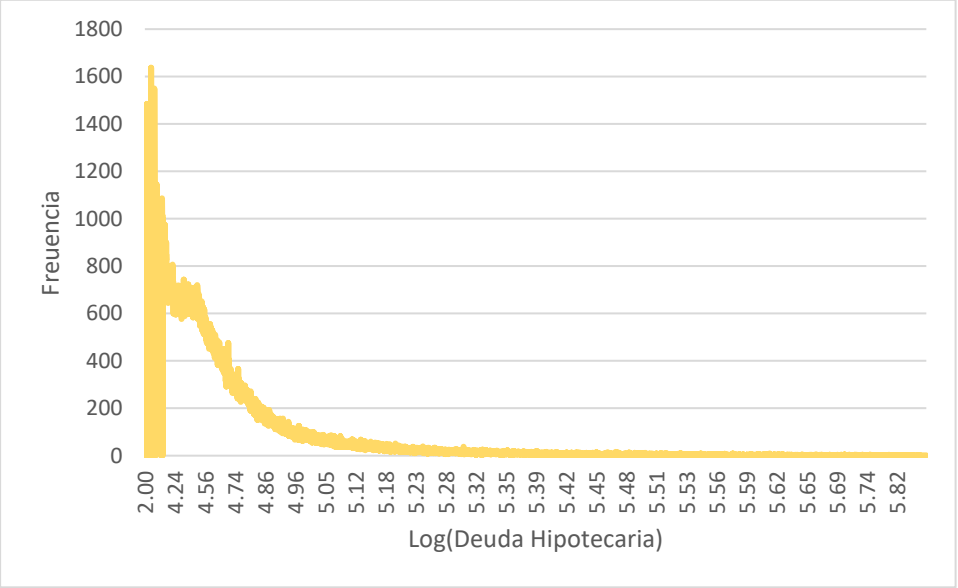
Fuente: Elaboración propia.

Como se puede ver, casi todos los clientes tienen deudas hipotecarias bajas o cercanas a 0.

Variable SBIF deuda hipotecaria (MNCRDI).

Variable que representa la disponibilidad de línea de crédito según la Super Intendencia de Bancos e Instituciones Financieras. Se hizo con un muestreo de 10.000 clientes

Gráfico 23:Línea de crédito de los clientes

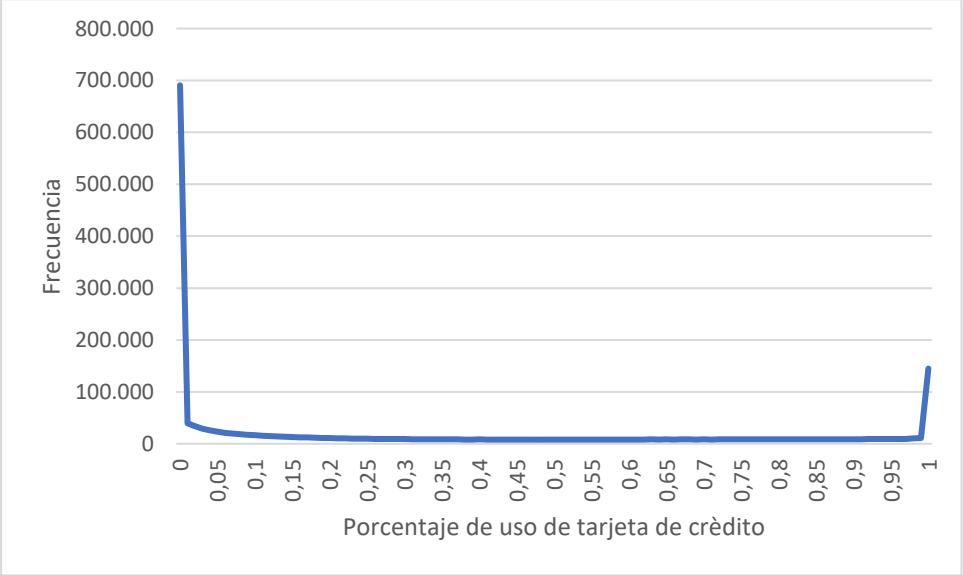


Fuente: Elaboración propia.

La deuda hipotecaria de los clientes se concentra en valores bajos

Porcentaje de uso de tarjeta (variable continua).
Se buscará el máximo de uso de la tarjeta mensual, y junto a esto el cómo usan los clientes la tarjeta de la empresa. Al ser un máximo por ID, se pudo realizar con la cartera de clientes activos completa.

Gráfico 24: Porcentaje de uso de la tarjeta.

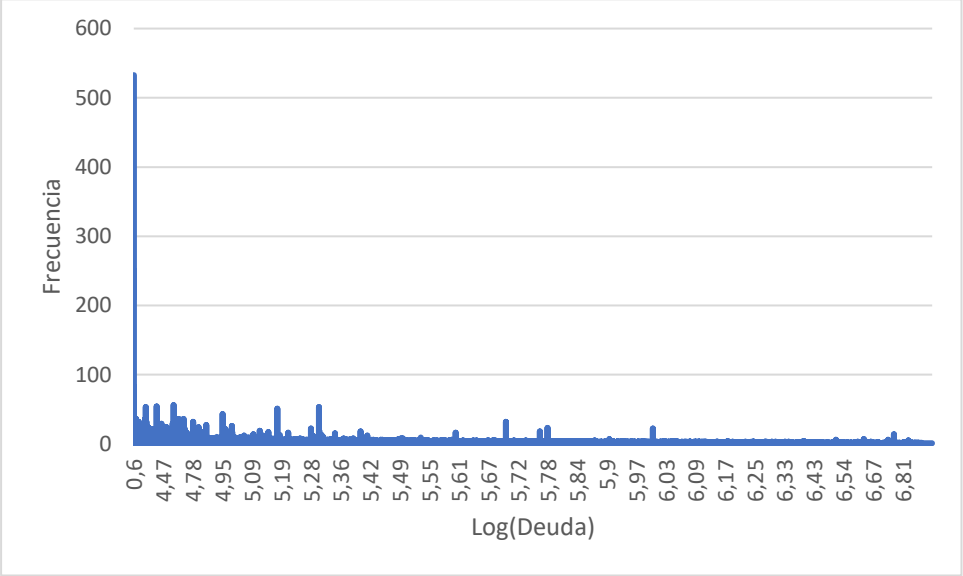


Fuente: Elaboración propia.

Este gráfico se hizo a partir de unos arreglos a la variable. Tenía valores negativos y además valores que sobrepasaban el 100%. Entonces, junto a la empresa se llegó al acuerdo de que todo valor negativo se le asignara un 0% y todo valor sobre 100% se le asignaba este último, por consiguiente, en las categorías 1 y 6 están todos aquellos valores que estaban fuera del rango, pero que no pueden ser eliminados. Estos errores ocurren debido a distintos factores: errores de imputación, al hacer devoluciones se restan porcentajes, entre otros.

Deuda del cliente con la empresa (variable continua).
Al igual que con las variables continuas SBIF, se le aplicó un logaritmo con tal de que la visualización sea más amena. Se hizo con un muestreo de 100.000 clientes.

Gráfico 25: Deuda de la cartera de clientes con la empresa.



Fuente: Elaboración propia.

.

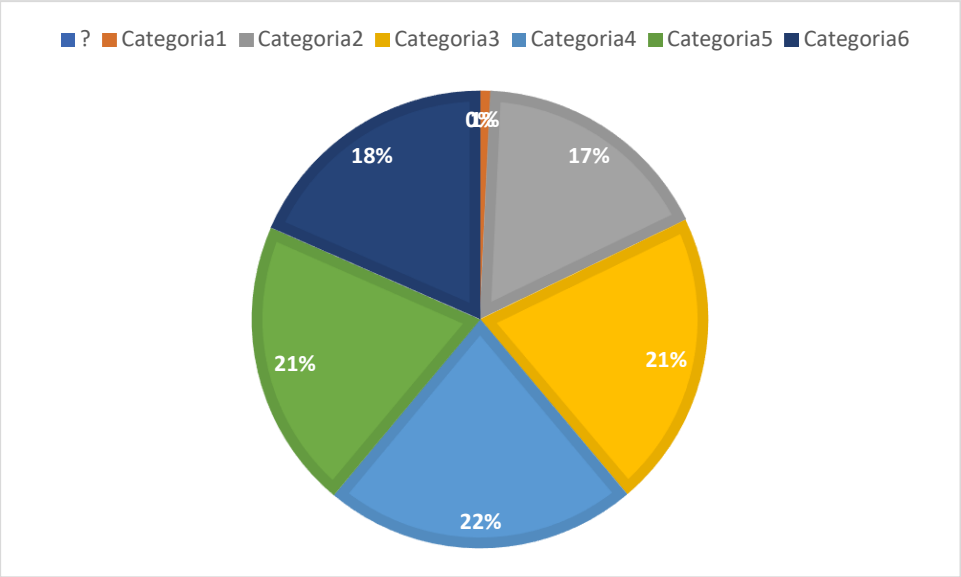
Edad Periodo

Si bien es una variable continua, se optó por elegir rangos etarios con el fin de que sea más fácil su ilustración e interpretación.

Quedando de la siguiente forma:
Categoría 1: Menores de 25 años.
Categoría 2: Mayores de 25 años, pero menores o iguales a 35 años.
Categoría 3: Mayores 35 años, pero menores o iguales a 45 años.
Categoría 4: Mayores de 45 años, pero menores o iguales a 55 años.

Categoría 5: Mayores de 55 años, pero menores o iguales a 65 años.
Categoría 6: Mayores a 66 años.

Gráfico 26:Número de clientes por categoría edad.



Fuente: Elaboración propia.

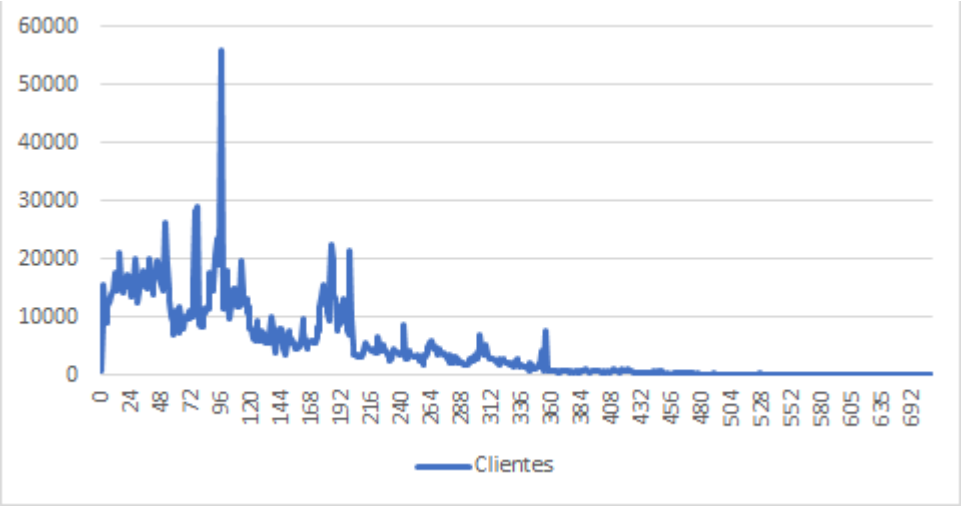
Se puede ver a simple vista que más del 65% de los clientes se distribuyen entre las categorías 3, 4 y 5, es decir, sus edades varían entre los 35 y 65 años. Hay que destacar también que alrededor de 22.000 clientes están en la categoría 1, representando casi un 1% de la cartera.

Variables dentro de la empresa:

Estas variables representan las acciones del cliente dentro de la empresa, como lo es:

Antigüedad del cliente
Desde cuando es cliente. Es una variable mensual que se actualiza todos los meses (variable continua).

Gráfico 27: Máxima antigüedad del cliente medida en meses.



Fuente: Elaboración propia.

En este gráfico se dejó fuera a los missing values, pero representan el 4,45% de la cartera de clientes de la empresa. He de destacar que la distribución muestra clientes con más de 600 meses de antigüedad. Al consultar el porqué de esto, esta variable se crea una vez que el cliente ejecutó una compra, por lo que los clientes que compraron en los primeros años de vida de Cencosud están considerados en la variable. Los números más repetidos de máxima antigüedad (es decir, su antigüedad actual), son 96 meses, que corresponden a aproximadamente 8 años de antigüedad en el retail.

Modelos homogéneos

Target: 1 mes.

Cabe destacar, que los datos totales en el testeo disminuyeron debido a que a medida que pasa el tiempo, empiezan a faltar meses disponibles para predecir (por ejemplo, no existe el 4to mes de predicción de noviembre de 2019, que vendría siendo marzo 2020, dato que no se tiene por lo que son NA's).

La matriz de confusión nos ayudará para evaluar cómo el modelo está separando cuando se le aplica los de testeo. Además, se compara con la matriz de los datos de entrenamiento, para que a simple vista logre verse que no se está ante un proceso de overfitting.

Tabla 15. Matriz de confusión datos de testeo, target 1 mes.

	Reales		
Predichos		0	1
	0	188.467	37.776

	1	39.346	222.471

Tabla 16. Matriz de confusión datos de entrenamiento, target 1 mes.

	Reales		
Predichos		0	1
	0	181.566	40.334
	1	29.243	178.366

Se puede concluir de esta comparación que el modelo logra predecir bien para los datos pertenecientes al 2019, y también logra calibrar de manera equitativa con los de entrenamiento.

Hecho esto, se muestran los coeficientes de la regresión logística:

Tabla 17. Tabla de coeficientes regresión logística, target 1 mes

	Variable Dependiente:
	Ingreso
(Intercept)	-4.712^{***}
	(0.225)
SBIF_DeudaVigente	-3.697
	(1.993)
SBIF_NumeroDeInstituciones	4.176^{***}
	(0.058)
SBIF_DeudaHipotecaria	-1.789^{***}
	(0.410)
SBIF_LineaDeCredito	-5.0562^{***}
	(0.567)
Edad	0.551^{***}
	(0.034)
MesAntiguedad	0.205^{***}
	(0.049)
PorcentajeUsoTarjeta	5.866^{***}
	(0.053)
Deuda	14.240^{***}
	(1.910)
TomaSegurosTCC	4.593^{***}
	(0.109)
TomaSegurosTCA	3.334^{***}
	(0.066)
TomaSAE	-0.902^{***}
	(0.223)
TomaAE	0.4773^{**}
	(0.147)
Enero	-0.137^{***}
	(0.0247)
Abril	0.078^{**}
	(0.024)
Mayo	0.061^{*}
	(0.024)
Junio	0.029
	(0.024)
Julio	-0.016
	(0.024)
Agosto	-0.053^{*}
	(0.024)
Septiembre	-0.076^{**}
	(0.024)
Octubre	0.059^{*}
	(0.024)

Noviembre	0.091^{***}
	(0.024)
PagoEnIntereses	2.080^{***}
	(0.087)
PagoMixto	1.707^{***}
	(0.098)
PagoEnContado	0.513^{***}
	(0.092)
Genero_M	-0.095^{***}
	(0.011)
region_1	0.006
	(0.049)
region_2	0.007
	(0.033)
region_3	0.057
	(0.047)
region_4	0.082^{*}
	(0.032)
region_5	-0.064^{**}
	(0.023)
region_6	-0.003
	(0.031)
region_7	-0.087^{**}
	(0.030)
region_8	-0.001
	(0.023)
region_9	$-0.056.$
	(0.031)
region_10	-0.014
	(0.033)
region_11	-0.100
	(0.143)
region_12	$-0.142.$
	(0.079)
region_14	-0.261^{***}
	(0.062)
region_15	0.080.
	(0.046)

RM_CentroNorte	−0.036
	(0.024)
RM_Occidente	0.003
	(0.025)
RM_Oriente	−0.088***
	(0.026)
RM_Periferica	0.045
	(0.044)
RM_Sur	−0.003
	(0.025)
RecenciaRetail1	0.118*
	(0.060)
RecenciaRetail2	0.231***
	(0.063)
RecenciaRetail3	0.272***
	(0.064)
RecenciaRetail4	0.271***
	(0.063)
RecenciaRetail5	0.445***
	(0.066)
RecenciaRetail6	0.472***
	(0.069)
RecenciaRetail7	0.322***
	(0.069)
RecenciaRetail8	0.297***
	(0.070)
RecenciaRetail9	0.144*
	(0.072)
RecenciaRetail10	0.166*
	(0.074)
RecenciaRetail11	0.219**
	(0.074)
RecenciaRetail12	0.104
	(0.075)

RecenciaTarjeta1	2.831***
	(0.109)
RecenciaTarjeta2	1.837***
	(0.110)
RecenciaTarjeta3	1.529***
	(0.111)
RecenciaTarjeta4	2.039***
	(0.062)
RecenciaTarjeta5	1.652***
	(0.066)
RecenciaTarjeta6	1.316***
	(0.069)
RecenciaTarjeta7	1.288***
	(0.069)
RecenciaTarjeta8	1.188***
	(0.070)
RecenciaTarjeta9	1.199***
	(0.072)
RecenciaTarjeta10	0.971***
	(0.075)
RecenciaTarjeta11	0.913***
	(0.074)
RecenciaTarjeta12	0.865***
	(0.076)
Note:	*p<0.1; **p<0.05; ***p<0.01

A continuación, tal como se hizo anteriormente, se muestra la matriz de confusión del WoE para ver su comportamiento:

Tabla 18. Matriz de confusión datos de testeo, target 1 mes WoE.

	Reales		
Predichos		0	1
	0	197.238	30.574
	1	38.728	221.510

Tabla 19. Matriz de confusión datos de entrenamiento, target 1 mes WoE.

	Reales		
Predichos		0	1
	0	179.589	31.219
	1	39.645	179.025

Se entregan también los coeficientes de la regresión logística con preprocesamiento WoE:

Tabla 20. Tabla de coeficientes regresión con WoE, target 1 mes.

	Variable Dependiente:
	Ingresos
(Intercept)	−0.091***
	(0.011)
Woe_SBIF_NINSDD_UM	0.088***
	(0.006)
Woe_SBIF_DDIRVG_UM	0.198***
	(0.007)
Woe_SBIF_MNCRDI_UM	0.369***
	(0.021)
Woe_MesAntiguedadCliente	0.359***
	(0.026)
Woe_EdadPeriodo	−0.306**
	(0.115)
Woe_SEXO	0.754***
	(0.080)
Woe_region_zona	0.508***
	(0.066)
Woe_RECENCIA_RETAIL	0.042***
	(0.012)
Woe_RECENCIA_TARJETA	0.374***
	(0.012)
Woe_porc_uso_adj	0.274***
	(0.005)
Woe_Deuda	0.602***
	(0.003)
Woe_MR_FORMA_TRX_U3M	0.125***
	(0.005)
Woe_mes	2.789***
	(0.280)
Woe_mark_trx_AE	0.237***
	(0.027)
Woe_mark_trx_SAE	0.007
	(0.047)
Woe_mark_trx_seguros_TCC	0.773***
	(0.025)
Woe_mark_trx_seguros_TCA	0.516***
	(0.015)
Note:	*p<0.1; **p<0.05; ***p<0.01

.

Target: 2 meses.

Matrices de confusión

Matriz de confusión de prueba

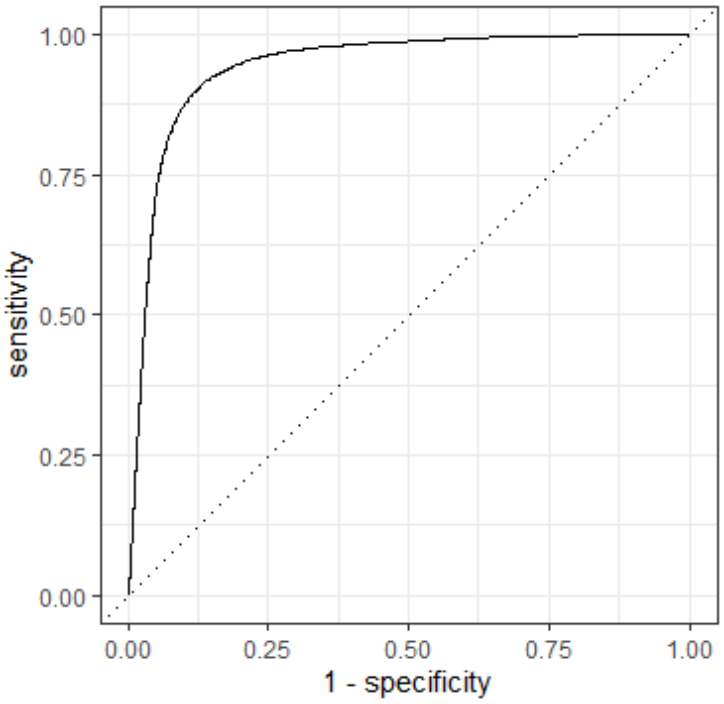
		Truth	
		0	1
Prediction	0	172.995	38.563
	1	40.681	198.019

Matriz de confusión train set

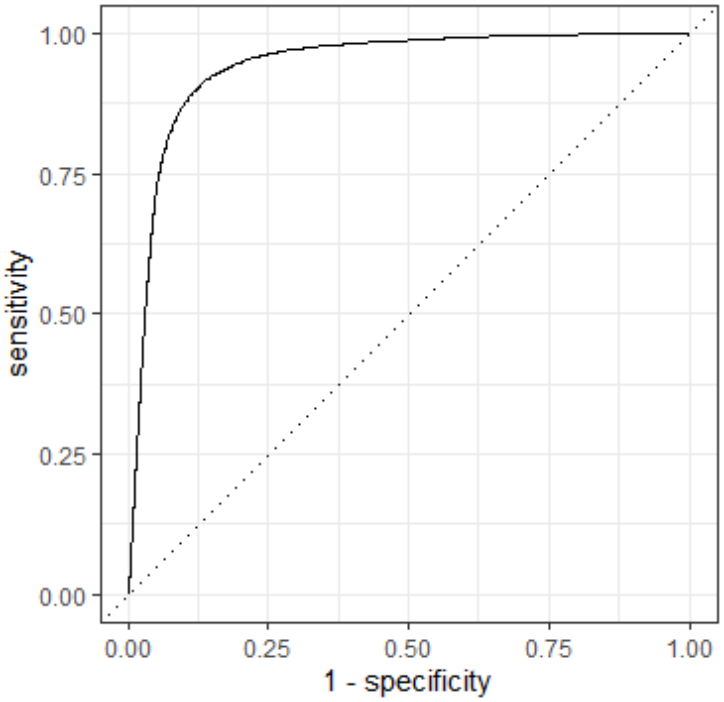
		Truth	
		0	1
Prediction	0	178.824	45.186
	1	35.137	170.396

ROC y AUC

Test: AUC: 0,862



Train: AUC: 0,859


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity

Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-8.687^{***}
	(0.209)
DDIRVG_norm	-2.682
	(2.351)
NINSDD_norm	3.577^{***}
	(0.049)
DDAHIP_norm	-1.601^{***}
	(0.196)
MNCRDI_norm	-6.685^{***}
	(0.703)
EdadPeriodo_norm	0.510^{***}
	(0.028)
MesAntiguedadCliente$_norm$	0.247^{***}
	(0.040)
Porc_uso_norm	2.426^{***}
	(0.025)
Deuda_norm	23.086^{***}
	(0.819)
mark_trx_seguros_TCC	3.352^{***}
	(0.059)
mark_trx_seguros_TCA	2.511^{***}
	(0.045)
mark_trx_SAE	0.479^{**}
	(0.174)
mark_trx_AE	0.621^{***}
	(0.082)
Enero	0.018
	(0.021)
Febrero	0.137^{***}
	(0.213)
Abril	0.102^{***}
	(0.021)
Mayo	0.059^{**}
	(0.021)
Junio	0.019
	(0.021)
Julio	0.068^{**}
	(0.021)

Agosto	0.170^{***}
	(0.021)
Septiembre	0.131^{***}
	(0.021)
Octubre	0.096^{***}
	(0.021)
Noviembre	0.011
	(0.021)
Forma_Interes	1.757^{***}
	(0.054)
Forma_Mixto	1.928^{***}
	(0.061)
Forma_Contado	1.071^{***}
	(0.057)
Sexo_M	-0.122^{***}
	(0.009)
Region_1	0.072.
	(0.040)
Region_2	0.044
	(0.028)
Region_3	0.145^{***}
	(0.038)
Region_4	0.091^{***}
	(0.027)
Region_5	-0.057^{**}
	(0.019)
Region_6	$-0.045.$
	(0.026)
Region_7	-0.085^{***}
	(0.025)
Region_8	-0.010
	(0.019)
Region_9	-0.012
	(0.026)
Region_10	-0.011
	(0.028)
Region_11	$-0.208.$
	(0.123)
Region_12	-0.052
	(0.066)
Region_14	$-0.087.$
	(0.051)
Region_15	0.072.
	(0.038)

RM_CentroNorte	-0.042^{*}
	(0.020)
RM_Occidente	0.013
	(0.021)
RM_Oriente	-0.110^{***}
	(0.022)
RM_Periferica	0.116^{**}
	(0.037)
RM_Sur	0.031
	(0.020)
RR_1	0.102^{*}
	(0.045)
RR_2	0.294^{***}
	(0.047)
RR_3	0.466^{***}
	(0.050)
RR_4	0.460^{***}
	(0.051)
RR_5	0.397^{***}
	(0.053)
RR_6	0.381^{***}
	(0.054)
RR_7	0.266^{***}
	(0.056)
RR_8	0.251^{***}
	(0.057)
RR_9	0.226^{***}
	(0.058)
RR_10	0.165^{**}
	(0.059)
RR_11	0.175^{**}
	(0.061)
RR_12	0.060
	(0.062)
RT_1	1.973^{***}
	(0.071)
RT_2	1.090^{***}
	(0.073)
RT_3	0.655^{***}
	(0.074)

RT_4	1.694*** (0.051)
RT_5	1.511*** (0.053)
RT_6	1.336*** (0.054)
RT_7	1.289*** (0.056)
RT_8	1.208*** (0.058)
RT_9	1.041*** (0.058)
RT_10	0.963*** (0.060)
RT_11	0.781*** (0.062)
RT_12	0.788*** (0.063)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	212.297	42.823
Obs. Pos	39.170	193.760

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	201.289	34.159
Obs. Pos	41.895	152.334

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.172^{***}
	(0.009)
Woe_SBIF_NINSDD_UM	0.083^{***}
	(0.004)
Woe_SBIF_DDIRVG_UM	-0.204^{***}
	(0.006)
Woe_SBIF_MNCRDI_UM	-0.533^{***}
	(0.016)
Woe_MesAntiguedadCliente	-0.251^{***}
	(0.021)
Woe_EdadPeriodo	0.204^{*}
	(0.088)
Woe_SEXO	-0.994^{***}
	(0.061)
Woe_region_zona	-0.588^{***}
	(0.053)
Woe_RECENCIA_RETAIL	-0.058^{***}
	(0.010)
Woe_RECENCIA_TARJETA	-0.395^{***}
	(0.010)
Woe_porc_uso_adj	-0.269^{***}
	(0.005)
Woe_Deuda	-0.499^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	-0.129^{***}
	(0.005)
Woe_mes	-1.224^{***}
	(0.181)
Woe_mark_trx_AE	-0.287^{***}
	(0.019)
Woe_mark_trx_SAE	-0.281^{***}
	(0.041)
Woe_mark_trx_seguros_TCC	-0.770^{***}
	(0.018)
Woe_mark_trx_seguros_TCA	-0.517^{***}
	(0.012)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 3 meses.

Matrices de confusión

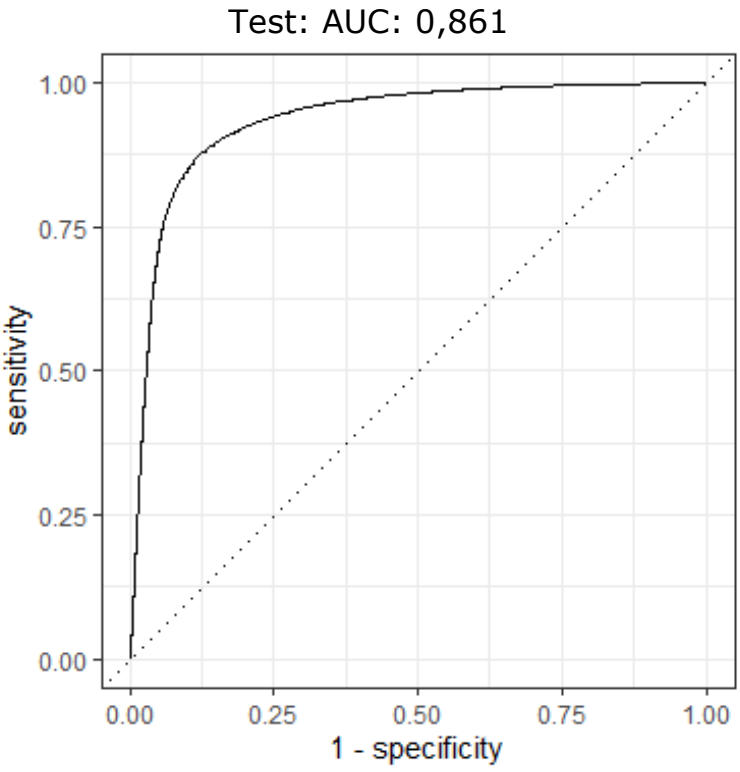
Matriz de confusión de prueba

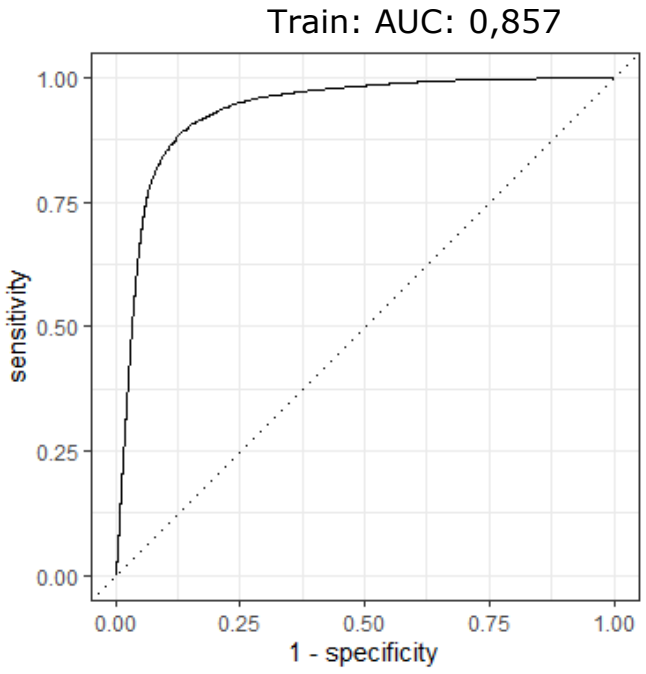
		Truth	
		0	1
Prediction	0	159.415	37.069
	1	40.138	176.251

Matriz de confusión train set

		Truth	
		0	1
Prediction	0	196.100	49.746
	1	40.984	185.722

ROC y AUC





Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-8.687^{***}
	(0.209)
DDIRVG_norm	-2.682
	(2.351)
NINSDD_norm	3.577^{***}
	(0.049)
DDAHIP_norm	-1.601^{***}
	(0.196)
MNCRDI_norm	-6.685^{***}
	(0.703)
EdadPeriodo_norm	0.510^{***}
	(0.028)
MesAntiguedadCliente$_norm$	0.247^{***}
	(0.040)
Porc_uso_norm	2.426^{***}
	(0.025)
Deuda_norm	23.086^{***}
	(0.819)
mark_trx_seguros_TCC	3.352^{***}
	(0.059)
mark_trx_seguros_TCA	2.511^{***}
	(0.045)
mark_trx_SAE	0.479^{**}
	(0.174)
mark_trx_AE	0.621^{***}
	(0.082)
Enero	0.018
	(0.021)
Febrero	0.137^{***}
	(0.213)
Abril	0.102^{***}
	(0.021)
Mayo	0.059^{**}
	(0.021)
Junio	0.019
	(0.021)
Julio	0.068^{**}
	(0.021)

Agosto	0.170^{***}
	(0.021)
Septiembre	0.131^{***}
	(0.021)
Octubre	0.096^{***}
	(0.021)
Noviembre	0.011
	(0.021)
Forma_Interes	1.757^{***}
	(0.054)
Forma_Mixto	1.928^{***}
	(0.061)
Forma_Contado	1.071^{***}
	(0.057)
Sexo_M	-0.122^{***}
	(0.009)
Region_1	0.072.
	(0.040)
Region_2	0.044
	(0.028)
Region_3	0.145^{***}
	(0.038)
Region_4	0.091^{***}
	(0.027)
Region_5	-0.057^{**}
	(0.019)
Region_6	$-0.045.$
	(0.026)
Region_7	-0.085^{***}
	(0.025)
Region_8	-0.010
	(0.019)
Region_9	-0.012
	(0.026)
Region_10	-0.011
	(0.028)
Region_11	$-0.208.$
	(0.123)
Region_12	-0.052
	(0.066)
Region_14	$-0.087.$
	(0.051)
Region_15	0.072.
	(0.038)

RM_CentroNorte	-0.042^{*}
	(0.020)
RM_Occidente	0.013
	(0.021)
RM_Oriente	-0.110^{***}
	(0.022)
RM_Periferica	0.116^{**}
	(0.037)
RM_Sur	0.031
	(0.020)
RR_1	0.102^{*}
	(0.045)
RR_2	0.294^{***}
	(0.047)
RR_3	0.466^{***}
	(0.050)
RR_4	0.460^{***}
	(0.051)
RR_5	0.397^{***}
	(0.053)
RR_6	0.381^{***}
	(0.054)
RR_7	0.266^{***}
	(0.056)
RR_8	0.251^{***}
	(0.057)
RR_9	0.226^{***}
	(0.058)
RR_10	0.165^{**}
	(0.059)
RR_11	0.175^{**}
	(0.061)
RR_12	0.060
	(0.062)
RT_1	1.973^{***}
	(0.071)
RT_2	1.090^{***}
	(0.073)
RT_3	0.655^{***}
	(0.074)

RT_4	1.694*** (0.051)
RT_5	1.511*** (0.053)
RT_6	1.336*** (0.054)
RT_7	1.289*** (0.056)
RT_8	1.208*** (0.058)
RT_9	1.041*** (0.058)
RT_10	0.963*** (0.060)
RT_11	0.781*** (0.062)
RT_12	0.788*** (0.063)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	162.679	41.159
Obs. Pos	36.875	172.162

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	193.652	43.431
Obs. Post	51.079	184.389

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.172^{***}
	(0.009)
Woe_SBIF_NINSDD_UM	0.083^{***}
	(0.004)
Woe_SBIF_DDIRVG_UM	-0.204^{***}
	(0.006)
Woe_SBIF_MNCRDI_UM	-0.533^{***}
	(0.016)
Woe_MesAntiguedadCliente	-0.251^{***}
	(0.021)
Woe_EdadPeriodo	0.204^{*}
	(0.088)
Woe_SEXO	-0.994^{***}
	(0.061)
Woe_region_zona	-0.588^{***}
	(0.053)
Woe_RECENCIA_RETAIL	-0.058^{***}
	(0.010)
Woe_RECENCIA_TARJETA	-0.395^{***}
	(0.010)
Woe_porc_uso_adj	-0.269^{***}
	(0.005)
Woe_Deuda	-0.499^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	-0.129^{***}
	(0.005)
Woe_mes	-1.224^{***}
	(0.181)
Woe_mark_trx_AE	-0.287^{***}
	(0.019)
Woe_mark_trx_SAE	-0.281^{***}
	(0.041)
Woe_mark_trx_seguros_TCC	-0.770^{***}
	(0.018)
Woe_mark_trx_seguros_TCA	-0.517^{***}
	(0.012)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target 4 meses
Matrices de confusión target 4 meses.

Tabla 21. Matriz de confusión datos de testeo, target 4 meses.

	Reales		
Predichos		0	1
	0	144.203	36.717
	1	39.861	154.610

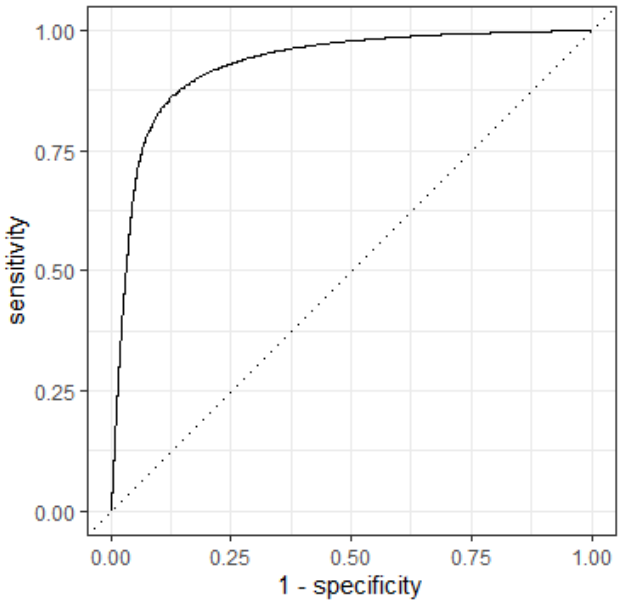
Esto entrega un accuracy de 86,7%.

Tabla 22. Matriz de confusión datos de entrenamiento, target 4 meses.

	Reales		
Predichos		0	1
	0	195.875	51.285
	1	43.253	182.272

Se procede a entregar las curvas ROC junto a su área bajo la curva.

Gráfico 28: Curva ROC para target a 4 meses, datos de testeo.

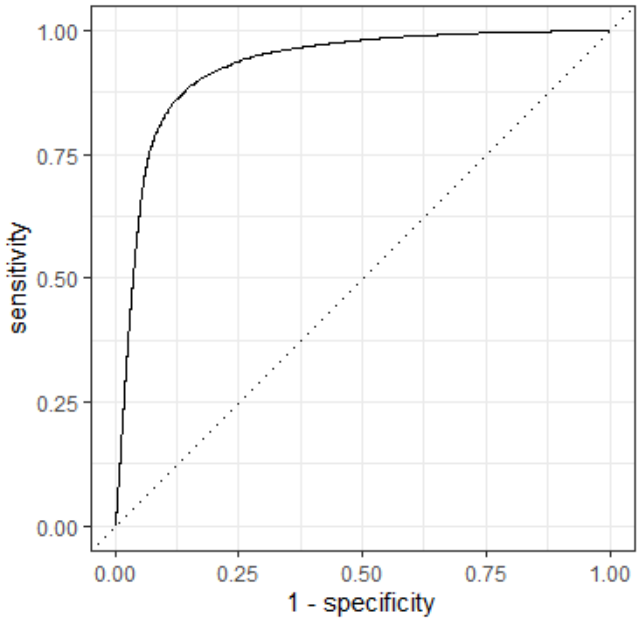


Fuente: Elaboración propia, a través de paquete parsnip en R.

El área bajo la curva de esta ROC es de 0,859, que indica que el modelo funciona bien con los datos que se le está entregando, al igual que en el primer modelo.

Para compararlo con el training set:

Gráfico 29:Curva ROC para target a 4 meses, datos de entrenamiento.



Fuente: Elaboración propia, a través de paquete parsnip en R.

El área bajo la curva es de 0,857. El modelo homogéneo para todos los periodos funcionó en buenos términos para esta variable independiente que sirve para predecir el periodo t+4.

Tabla 23. Tabla de coeficientes regresión logística

	Variable Dependiente:
	Ingresos
(Intercept)	-8.462^{***}
	(0.171)
DDIRVG_norm	18.419^{**}
	(6.670)
NINSDD_norm	3.000^{***}
	(0.044)
DDAHIP_norm	-4.203^{***}
	(0.497)
MNCRDI_norm	-8.249^{***}
	(0.786)
EdadPeriodo_norm	0.416^{***}
	(0.025)
MesAntiguedadCliente_norm	0.296^{***}
	(0.038)
Porc_uso_norm	1.100^{***}
	(0.018)
Deuda_norm	13.526^{***}
	(0.391)
mark_trx_seguros_TCC	2.303^{***}
	(0.040)
mark_trx_seguros_TCA	1.794^{***}
	(0.032)
mark_trx_SAE	0.576^{***}
	(0.110)
mark_trx_AE	0.311^{***}
	(0.047)
Enero	-0.050^{*}
	(0.019)
Febrero	0.110^{***}
	(0.019)
Abril	0.159^{***}
	(0.019)
Mayo	0.097^{***}
	(0.019)
Junio	0.068^{***}
	(0.019)
Julio	0.062^{**}
	(0.019)

Agosto	0.083^{***}
	(0.019)
Septiembre	0.086^{***}
	(0.019)
Octubre	0.120^{***}
	(0.019)
Noviembre	0.048^{*}
	(0.019)
Forma_Interes	1.661^{***}
	(0.040)
Forma_Mixto	1.885^{***}
	(0.046)
Forma_Contado	1.257^{***}
	(0.044)
Sexo_M	-0.121^{***}
	(0.008)
Region_1	0.055
	(0.036)
Region_2	0.020
	(0.026)
Region_3	0.069^{*}
	(0.035)
Region_4	0.063^{*}
	(0.024)
Region_5	-0.023
	(0.018)
Region_6	-0.005
	(0.024)
Region_7	-0.054^{*}
	(0.023)
Region_8	-0.004
	(0.017)
Region_9	-0.048^{*}
	(0.023)
Region_10	0.038
	(0.025)
Region_11	-0.135
	(0.113)
Region_12	0.037
	(0.059)
Region_14	-0.055
	(0.046)
Region_15	0.149^{***}
	(0.035)

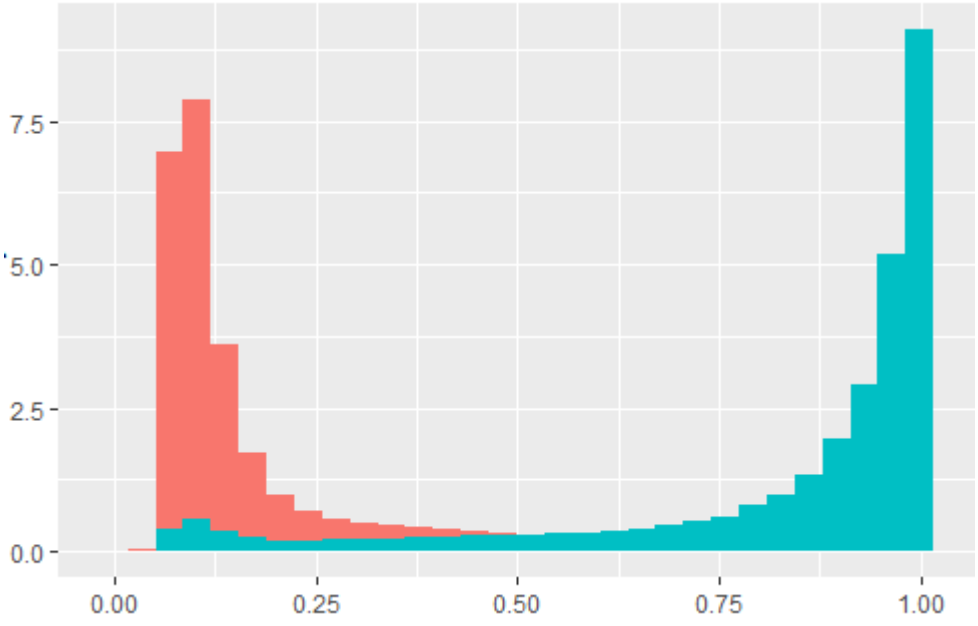
RM_CentroNorte	$-0.033.$ (0.018)
RM_Occidente	-0.000 (0.019)
RM_Oriente	-0.193^{***} (0.021)
RM_Periferica	0.096^{**} (0.033)
RM_Sur	0.045^{*} (0.019)
RR_1	0.429^{***} (0.035)
RR_2	0.410^{***} (0.038)
RR_3	0.406^{***} (0.041)
RR_4	0.416^{***} (0.043)
RR_5	0.384^{***} (0.045)
RR_6	0.209^{***} (0.046)
RR_7	0.260^{***} (0.049)
RR_8	0.234^{***} (0.050)
RR_9	0.190^{***} (0.053)
RR_10	$0.102.$ (0.054)
RR_11	0.126^{*} (0.056)
RR_12	-0.037 (0.057)
RT_1	1.324^{***} (0.054)
RT_2	0.664^{***} (0.056)
RT_3	0.438^{***} (0.058)
RT_4	1.6554^{***} (0.044)

RT_5	1.348*** (0.046)
RT_6	1.363*** (0.047)
RT_7	1.164*** (0.050)
RT_8	1.032*** (0.051)
RT_9	0.970*** (0.054)
RT_10	0.863*** (0.055)
RT_11	0.725*** (0.057)
RT_12	0.832*** (0.058)
Note:	*p<0.1; **p<0.05; ***p<0.01

A la predicción de 4 meses siguientes, en comparación al primer modelo, se le ve que no sigue la misma línea de la recencia en la tarjeta, pero aun así se mantiene en formato positivo. Se sigue manteniendo el peso del sector Oriente respecto a la región Metropolitana, pero con un efecto negativo, el género masculino tiene un factor negativo con lo que respecta a la probabilidad de compra y tener deuda con la empresa hace un efecto positivo en la probabilidad de compra.

Al igual que el modelo a un mes de predicción, se entregan los resultados normalizados para ver cómo está separando el modelo las probabilidades:

Gráfico 30: Separación del modelo en target a 4 meses.



Fuente: Elaboración propia.

Se demuestra entonces que el modelo sabe diferenciar las probabilidades y logra categorizarlas como corresponde.

Se procede a la comparación de la matriz de confusión del WoE para ver su comportamiento junto a sus coeficientes:

Matrices de confusión WoE

Tabla 24 Matriz de confusión datos de testeo, target 4 mes WoE.

	Reales		
Predichos		0	1
	0	149.116	34.948
	1	38.252	153.075

Tabla 25. Matriz de confusión datos de entrenamiento, target 4 mes WoE.

	Reales		
Predichos		0	1
	0	193.914	51.215
	1	45.214	182.342

A continuación, la tabla de los coeficientes de la regresión con WoE:

Tabla 26. Tabla de coeficientes de regresión WoE, target 4 meses.

	Variable Dependiente:
	Ingresos
(Intercept)	-0.146^{***}
	(0.004)
Woe_SBIF_NINSDD_UM	0.101^{***}
	(0.006)
Woe_SBIF_DDIRVG_UM	0.182^{***}
	(0.007)
Woe_SBIF_MNCRDI_UM	0.548^{***}
	(0.013)
Woe_MesAntiguedadCliente	0.257^{***}
	(0.019)
Woe_EdadPeriodo	0.467^{***}
	(0.072)
Woe_SEXO	0.964^{***}
	(0.052)
Woe_region_zona	0.606^{***}
	(0.051)
Woe_RECENCIA_RETAIL	0.111^{***}
	(0.009)
Woe_RECENCIA_TARJETA	0.353^{***}
	(0.010)
Woe_porc_uso_adj	0.200^{***}
	(0.005)
Woe_Deuda	0.457^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	0.154^{***}
	(0.005)
Woe_mark_trx_AE	0.236^{***}
	(0.015)
Woe_mark_trx_SAE	0.258^{***}
	(0.033)
Woe_mark_trx_seguros_TCC	0.710^{***}
	(0.016)
Woe_mark_trx_seguros_TCA	0.471^{***}
	(0.010)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 5 meses.

Matrices de confusión

Matriz de confusión de prueba

		Truth	
		0	1
Prediction	0	133.443	32.523
	1	35.561	135.528

Matriz de confusión train set

		Truth	
		0	1
Prediction	0	196.405	52.951

	1	45.543	178.633

ROC y AUC
Test: AUC: 0,858

Train:AUC: 0,855

Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-7.094^{***}
	(0.258)
DDIRVG_norm	-2.704
	(1.706)
NINSDD_norm	3.585^{***}
	(0.050)
DDAHIP_norm	-2.060^{**}
	(0.682)
MNCRDI_norm	-6.993^{***}
	(0.753)
EdadPeriodo_norm	0.631^{***}
	(0.032)
MesAntiguedadCliente_norm	0.228^{***}
	(0.046)
Porc_uso_norm	4.026^{***}
	(0.037)
Deuda_norm	11.983^{***}
	(0.760)
mark_trx_seguros_TCC	3.956^{***}
	(0.081)
mark_trx_seguros_TCA	3.136^{***}
	(0.062)
mark_trx_SAE	-0.143
	(0.211)
mark_trx_AE	0.466^{***}
	(0.111)
Febrero	0.058^{*}
	(0.023)
Abril	0.170^{***}
	(0.023)
Mayo	0.151^{***}
	(0.022)
Junio	0.121^{***}
	(0.023)
Julio	0.091^{***}
	(0.023)

Agosto	0.073**
	(0.023)
Septiembre	0.104***
	(0.023)
Octubre	0.216***
	(0.022)
Noviembre	0.128***
	(0.023)
Forma_Interes	1.947***
	(0.070)
Forma_Mixto	1.833***
	(0.079)
Forma_Contado	0.759***
	(0.074)
Sexo_M	−0.114***
	(0.010)
Region_1	0.011
	(0.046)
Region_2	0.059.
	(0.032)
Region_3	0.100*
	(0.043)
Region_4	0.040
	(0.030)
Region_5	−0.102***
	(0.022)
Region_6	−0.010
	(0.029)
Region_7	−0.088**
	(0.028)
Region_8	−0.040.
	(0.022)
Region_9	−0.102***
	(0.029)
Region_10	−0.002
	(0.031)
Region_11	−0.342*
	(0.140)
Region_12	−0.207**
	(0.075)
Region_14	−0.226***
	(0.058)
Region_15	0.141**
	(0.044)

RM_CentroNorte	-0.069^{**}
	(0.023)
RM_Occidente	-0.020
	(0.024)
RM_Oriente	-0.092^{***}
	(0.025)
RM_Periferica	0.097^{*}
	(0.041)
RM_Sur	-0.010
	(0.023)
RR_1	0.125^{*}
	(0.545)
RR_2	0.288^{***}
	(0.056)
RR_3	0.245^{***}
	(0.059)
RR_4	0.356^{***}
	(0.059)
RR_5	0.543^{***}
	(0.063)
RR_6	0.524^{***}
	(0.064)
RR_7	0.252^{***}
	(0.065)
RR_8	0.186^{**}
	(0.066)
RR_9	0.240^{***}
	(0.067)
RR_10	0.261^{***}
	(0.069)
RR_11	0.321
	(0.067)
RR_12	0.119.
	(0.070)
RT_1	2.431^{***}
	(0.091)
RT_2	1.477^{***}
	(0.091)
RT_3	1.213^{***}
	(0.093)
RT_4	1.896^{***}
	(0.058)

RT_5	1.446*** (0.062)
RT_6	1.222*** (0.064)
RT_7	1.312*** (0.065)
RT_8	1.232*** (0.066)
RT_9	1.067*** (0.068)
RT_10	0.941*** (0.069)
RT_11	1.058*** (0.068)
RT_12	0.773*** (0.071)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	125.109	29.036
Obs. Pos	29.907	116.671

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	195.126	46.824
Obs. Pos	51.198	180.385

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.267^{***}
	(0.069)
Woe_SBIF_NINSDD_UM	0.086^{***}
	(0.006)
Woe_SBIF_DDIRVG_UM	-0.210^{***}
	(0.007)
Woe_SBIF_MNCRDI_UM	-0.506^{***}
	(0.022)
Woe_MesAntiguedadCliente	-0.342^{***}
	(0.026)
Woe_EdadPeriodo	0.621^{***}
	(0.136)
Woe_SEXO	-0.996^{***}
	(0.082)
Woe_region_zona	-0.694^{***}
	(0.072)
Woe_RECENCIA_RETAIL	-0.049^{***}
	(0.014)
Woe_RECENCIA_TARJETA	-0.455^{***}
	(0.015)
Woe_porc_uso_adj	-0.502^{***}
	(0.008)
Woe_Deuda	-0.662^{***}
	(0.004)
Woe_MR_FORMA_TRX_U3M	-0.186^{***}
	(0.007)
Woe_mes	-2.896^{***}
	(0.342)
Woe_mark_trx_AE	-0.519^{***}
	(0.048)
Woe_mark_trx_SAE	-0.376^{***}
	(0.096)
Woe_mark_trx_seguros_TCC	-1.417^{***}
	(0.043)
Woe_mark_trx_seguros_TCA	-0.922^{***}
	(0.025)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 6 meses.

Matrices de confusión

Matriz de confusión de prueba

		Truth	
		0	1
Prediction	0	122.100	29.012
	1	32.044	117.575

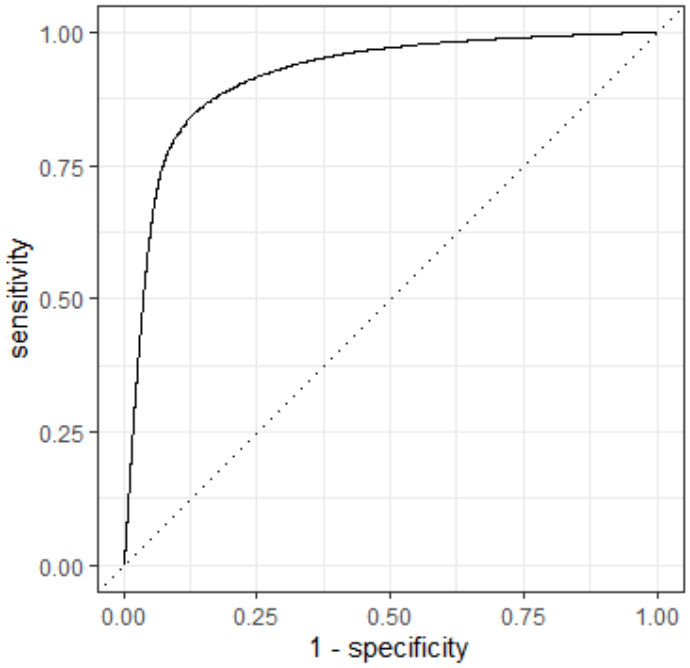
Matriz de confusión train set

		Truth

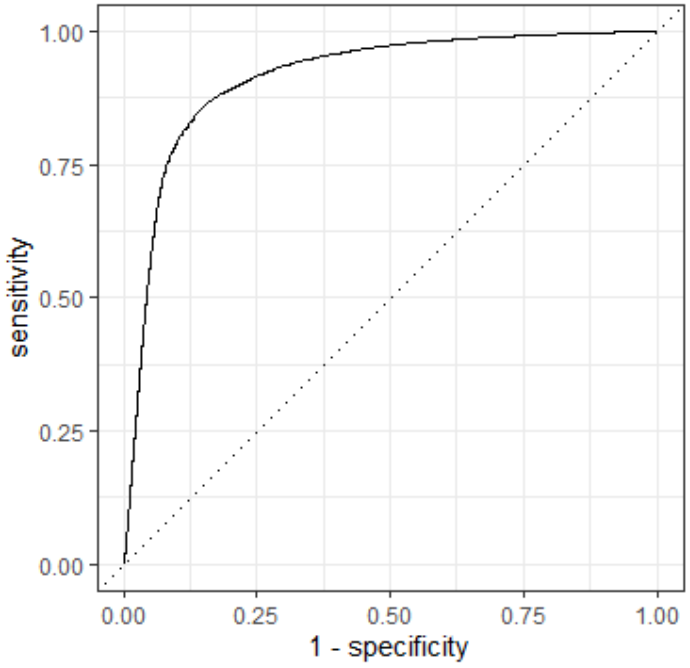
		0	1
Prediction	0	198.152	52.792
	1	46.040	175.887

ROC y AUC

Test: AUC: 0,857.



Train: AUC: 0,850



Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-7.457^{***}
	(0.221)
DDIRVG_norm	$-2.779.$
	(1.429)
NINSDD_norm	3.686^{***}
	(0.050)
DDAHIP_norm	-2.225^{***}
	(0.525)
MNCRDI_norm	-6.066^{***}
	(0.593)
EdadPeriodo_norm	0.572^{***}
	(0.029)
MesAntiguedadCliente_norm	0.281^{***}
	(0.042)
Porc_uso_norm	3.114^{***}
	(0.028)
Deuda_norm	12.081^{***}
	(0.588)
mark_trx_seguros_TCC	3.711^{***}
	(0.067)
mark_trx_seguros_TCA	2.821^{***}
	(0.051)
mark_trx_SAE	0.046
	(0.169)
mark_trx_AE	0.567^{***}
	(0.092)
Enero	$-0.037.$
	(0.022)
Febrero	0.136^{***}
	(0.022)
Abril	0.154^{***}
	(0.022)
Mayo	0.129^{***}
	(0.022)
Junio	0.087^{***}
	(0.022)
Julio	0.073^{**}
	(0.022)

Agosto	0.111^{***}
	(0.022)
Septiembre	0.149^{***}
	(0.022)
Octubre	0.230^{***}
	(0.021)
Noviembre	0.053^{*}
	(0.022)
Forma_Interes	1.918^{***}
	(0.059)
Forma_Mixto	2.078^{***}
	(0.066)
Forma_Contado	1.078^{***}
	(0.062)
Sexo_M	-0.098^{***}
	(0.009)
Region_1	0.003
	(0.041)
Region_2	0.013
	(0.029)
Region_3	0.001
	(0.040)
Region_4	0.078^{**}
	(0.027)
Region_5	-0.068^{***}
	(0.020)
Region_6	-0.072^{**}
	(0.027)
Region_7	-0.083^{**}
	(0.026)
Region_8	-0.044^{*}
	(0.020)
Region_9	$-0.051.$
	(0.027)
Region_10	-0.028
	(0.029)
Region_11	-0.140
	(0.122)
Region_12	-0.086
	(0.068)
Region_14	-0.219^{***}
	(0.053)
Region_15	0.123^{**}
	(0.040)

RM_CentroNorte	-0.090^{***}
	(0.021)
RM_Occidente	-0.019
	(0.022)
RM_Oriente	-0.131^{***}
	(0.023)
RM_Periferica	0.031
	(0.041)
RM_Sur	-0.006
	(0.021)
RR_1	0.253^{***}
	(0.047)
RR_2	0.319^{***}
	(0.050)
RR_3	0.428^{***}
	(0.053)
RR_4	0.520^{***}
	(0.054)
RR_5	0.561^{***}
	(0.056)
RR_6	0.311^{***}
	(0.056)
RR_7	0.319^{***}
	(0.058)
RR_8	0.193^{**}
	(0.059)
RR_9	0.366^{***}
	(0.062)
RR_10	0.197^{**}
	(0.062)
RR_11	0.161^{*}
	(0.063)
RR_12	0.076
	(0.065)
RT_1	1.878^{***}
	(0.077)
RT_2	1.074^{***}
	(0.078)
RT_3	0.752^{***}
	(0.080)
RT_4	1.684^{***}
	(0.054)

RT_5	1.443***
	(0.056)
RT_6	1.414***
	(0.056)
RT_7	1.309***
	(0.059)
RT_8	1.257***
	(0.060)
RT_9	0.999***
	(0.063)
RT_10	1.015***
	(0.062)
RT_11	0.921***
	(0.064)
RT_12	0.760***
	(0.066)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	123.906	30.238
Obs. Pos	31.109	115.469

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	196.376	47.815
Obs. Pos	51.015	177.664

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.140^{***}
	(0.009)
Woe_SBIF_NINSDD_UM	0.074^{***}
	(0.005)
Woe_SBIF_DDIRVG_UM	0.207^{***}
	(0.006)
Woe_SBIF_MNCRDI_UM	0.486^{***}
	(0.017)
Woe_MesAntiguedadCliente	0.289^{***}
	(0.022)
Woe_EdadPeriodo	-0.216^{*}
	(0.096)
Woe_SEXO	0.815^{***}
	(0.063)
Woe_region_zona	0.580^{***}
	(0.055)
Woe_RECENCIA_RETAIL	0.082^{***}
	(0.010)
Woe_RECENCIA_TARJETA	0.365^{***}
	(0.010)
Woe_porc_uso_adj	0.277^{***}
	(0.005)
Woe_Deuda	0.524^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	0.130^{***}
	(0.005)
Woe_mes	2.296^{***}
	(0.198)
Woe_mark_trx_AE	0.282^{***}
	(0.020)
Woe_mark_trx_SAE	0.171^{***}
	(0.040)
Woe_mark_trx_seguros_TCC	0.780^{***}
	(0.019)
Woe_mark_trx_seguros_TCA	0.534^{***}
	(0.013)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 7 meses.

Matrices de confusión

Matriz de confusión de prueba

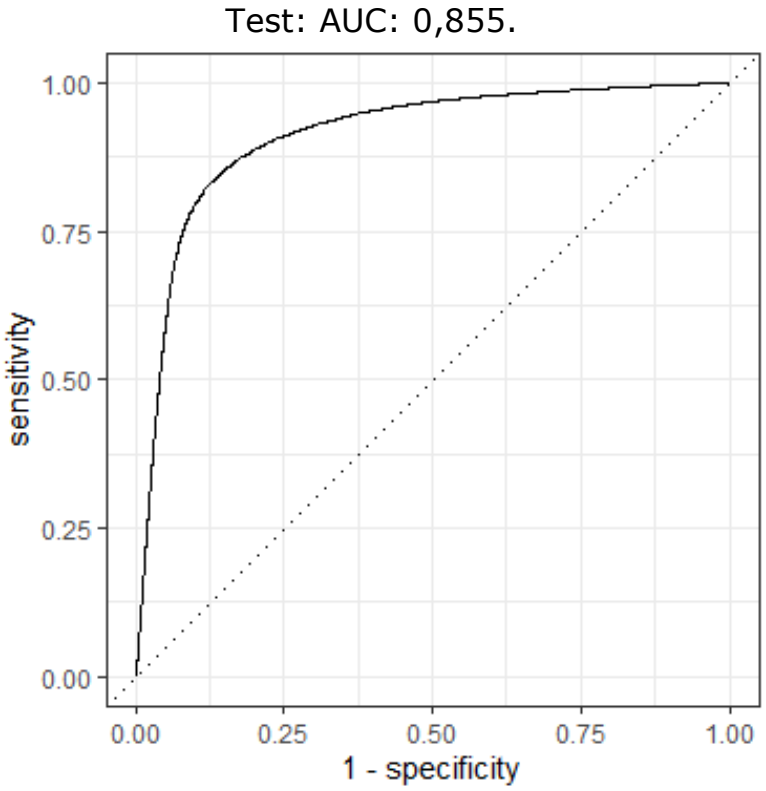
		Truth	
		0	1
Prediction	0	103.081	25.545
	1	28.191	106.591

Matriz de confusión train set

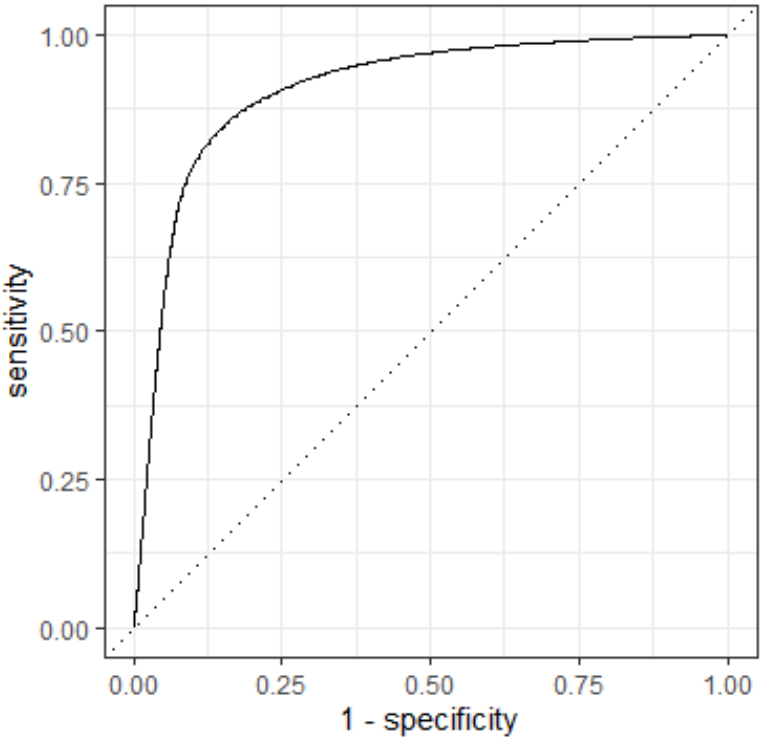
		Truth	
		0	1

Prediction	0	199.745	46.613
	1	52.895	174.602

ROC y AUC



Train: AUC: 0,842



Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-7.214^{***}
	(0.157)
DDIRVG_norm	-0.002
	(2.633)
NINSDD_norm	3.131^{***}
	(0.043)
DDAHIP_norm	-2.591^{***}
	(0.413)
MNCRDI_norm	-6.480^{***}
	(0.521)
EdadPeriodo_norm	0.534^{***}
	(0.028)
MesAntiguedadCliente_norm	0.248^{***}
	(0.038)
Porc_uso_norm	2.031^{***}
	(0.022)
Deuda_norm	15.273^{***}
	(0.538)
mark_trx_seguros_TCC	2.975^{***}
	(0.051)
mark_trx_seguros_TCA	2.298^{***}
	(0.041)
mark_trx_SAE	0.412^{**}
	(0.143)
mark_trx_AE	0.739^{***}
	(0.075)
Enero	0.034
	(0.020)
Febrero	0.106^{***}
	(0.020)
Abril	0.068^{**}
	(0.020)
Mayo	0.072^{***}
	(0.020)
Junio	0.126^{***}
	(0.020)
Julio	0.194^{**}
	(0.020)

Agosto	0.126^{***}
	(0.020)
Septiembre	0.035.
	(0.020)
Octubre	0.064^{**}
	(0.020)
Noviembre	0.032
	(0.020)
Forma_Interes	1.724^{***}
	(0.048)
Forma_Mixto	2.025^{***}
	(0.055)
Forma_Contado	1.200^{***}
	(0.052)
Sexo_M	-0.114^{***}
	(0.009)
Region_1	0.001
	(0.039)
Region_2	0.045.
	(0.027)
Region_3	0.130^{***}
	(0.037)
Region_4	0.128^{***}
	(0.026)
Region_5	-0.010
	(0.019)
Region_6	0.024
	(0.025)
Region_7	-0.033
	(0.024)
Region_8	-0.044^{*}
	(0.020)
Region_9	-0.033
	(0.025)
Region_10	0.066^{*}
	(0.027)
Region_11	0.029
	(0.115)
Region_12	-0.072
	(0.063)
Region_14	-0.062
	(0.050)
Region_15	0.126^{**}
	(0.038)

RM_CentroNorte	0.001
	(0.019)
RM_Occidente	0.057**
	(0.022)
RM_Oriente	−0.130***
	(0.022)
RM_Periferica	0.106**
	(0.036)
RM_Sur	0.049*
	(0.020)
RR_1	0.270***
	(0.041)
RR_2	0.418***
	(0.044)
RR_3	0.410***
	(0.047)
RR_4	0.514***
	(0.048)
RR_5	0.397***
	(0.050)
RR_6	0.378***
	(0.052)
RR_7	0.219***
	(0.053)
RR_8	0.208***
	(0.055)
RR_9	0.178**
	(0.056)
RR_10	0.224***
	(0.058)
RR_11	0.181**
	(0.059)
RR_12	0.063
	(0.061)
RT_1	1.643***
	(0.065)
RT_2	0.777***
	(0.066)
RT_3	0.568***
	(0.068)
RT_4	1.553***
	(0.048)

RT_5	1.458*** (0.050)
RT_6	1.313*** (0.052)
RT_7	1.340*** (0.054)
RT_8	1.092*** (0.055)
RT_9	1.016*** (0.057)
RT_10	0.899*** (0.059)
RT_11	0.778*** (0.060)
RT_12	0.693*** (0.061)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	111.281	27.316
Obs. Pos	27.474	97.377

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	195.110	51.249
Obs. Pos	53.000	174.497

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.183^{***}
	(0.008)
Woe_SBIF_NINSDD_UM	0.075^{***}
	(0.004)
Woe_SBIF_DDIRVG_UM	-0.204^{***}
	(0.006)
Woe_SBIF_MNCRDI_UM	-0.525^{***}
	(0.015)
Woe_MesAntiguedadCliente	-0.246^{***}
	(0.021)
Woe_EdadPeriodo	0.153.
	(0.086)
Woe_SEXO	-0.877^{***}
	(0.057)
Woe_region_zona	-0.610^{***}
	(0.055)
Woe_RECENCIA_RETAIL	-0.080^{***}
	(0.009)
Woe_RECENCIA_TARJETA	-0.371^{***}
	(0.010)
Woe_porc_uso_adj	-0.242^{***}
	(0.005)
Woe_Deuda	-0.493^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	-0.141^{***}
	(0.005)
Woe_mes	-1.306^{***}
	(0.168)
Woe_mark_trx_AE	-0.312^{***}
	(0.018)
Woe_mark_trx_SAE	-0.268^{***}
	(0.037)
Woe_mark_trx_seguros_TCC	-0.751^{***}
	(0.017)
Woe_mark_trx_seguros_TCA	-0.507^{***}
	(0.011)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 8 meses.

Matrices de confusión

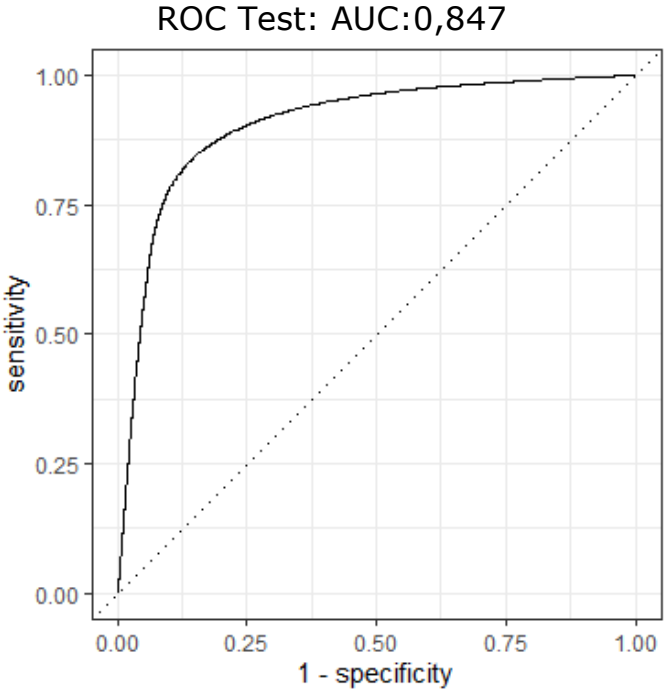
Matriz de confusión de prueba

		Truth	
		0	1
Prediction	0	87.429	21.603
	1	23.501	87.487

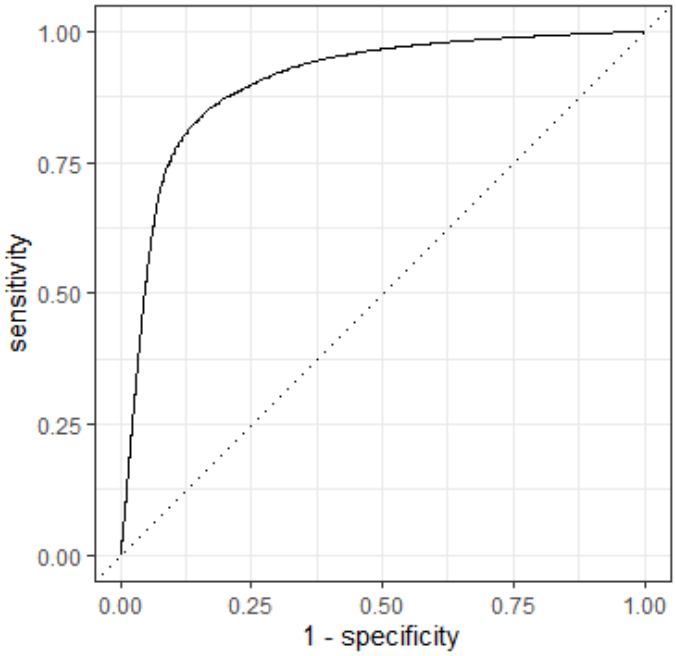
Matriz de confusión train set

		Truth	
		0	1
Prediction	0	201.140	52.356
	1	47.370	171.766

ROC y AUC



ROC Train: AUC: 0,838



Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-6.087^{***}
	(0.119)
DDIRVG_norm	-2.293
	(2.211)
NINSDD_norm	2.659^{***}
	(0.038)
DDAHIP_norm	-2.623^{***}
	(0.351)
MNCRDI_norm	-6.720^{***}
	(0.676)
EdadPeriodo_norm	0.506^{***}
	(0.027)
MesAntiguedadCliente_norm	0.327^{***}
	(0.039)
Porc_uso_norm	1.748^{***}
	(0.021)
Deuda_norm	13.546^{***}
	(0.468)
mark_trx_seguros_TCC	2.797^{***}
	(0.048)
mark_trx_seguros_TCA	2.087^{***}
	(0.038)
mark_trx_SAE	0.410^{**}
	(0.134)
mark_trx_AE	0.641^{***}
	(0.065)
Enero	-0.015
	(0.020)
Febrero	0.046^{*}
	(0.020)
Abril	0.060^{**}
	(0.020)
Mayo	0.116^{***}
	(0.020)
Junio	0.133^{***}
	(0.020)
Julio	0.138^{***}
	(0.020)

Agosto	0.031
	(0.020)
Septiembre	0.026
	(0.020)
Octubre	0.094***
	(0.020)
Noviembre	−0.005
	(0.020)
Forma_Interes	1.723***
	(0.045)
Forma_Mixto	1.977***
	(0.052)
Forma_Contado	1.265***
	(0.049)
Sexo_M	−0.118***
	(0.008)
Region_1	0.042
	(0.038)
Region_2	0.020
	(0.027)
Region_3	0.100**
	(0.036)
Region_4	0.048.
	(0.025)
Region_5	−0.031.
	(0.018)
Region_6	−0.020
	(0.018)
Region_7	−0.063**
	(0.024)
Region_8	−0.020
	(0.018)
Region_9	−0.056*
	(0.025)
Region_10	−0.017
	(0.026)
Region_11	−0.161
	(0.113)
Region_12	−0.022
	(0.063)
Region_14	−0.147**
	(0.049)
Region_15	0.118**
	(0.037)

RM_CentroNorte	-0.041^{*}
	(0.019)
RM_Occidente	0.017
	(0.020)
RM_Oriente	-0.109^{***}
	(0.022)
RM_Periferica	0.076^{*}
	(0.035)
RM_Sur	0.036.
	(0.020)
RR_1	0.199^{***}
	(0.040)
RR_2	0.356^{***}
	(0.042)
RR_3	0.360^{***}
	(0.046)
RR_4	0.343^{***}
	(0.047)
RR_5	0.444^{***}
	(0.049)
RR_6	0.312^{***}
	(0.050)
RR_7	0.182^{***}
	(0.052)
RR_8	0.289^{***}
	(0.054)
RR_9	0.070
	(0.055)
RR_10	0.145^{*}
	(0.057)
RR_11	$0.254^{*]*}$
	(0.058)
RR_12	0.115^{*}
	(0.058)
RT_1	1.633^{***}
	(0.062)
RT_2	0.795^{***}
	(0.063)
RT_3	0.519^{***}
	(0.065)
RT_4	1.686^{***}
	(0.047)

RT_5	1.406*** (0.049)
RT_6	1.341*** (0.051)
RT_7	1.322*** (0.052)
RT_8	1.044*** (0.054)
RT_9	1.103*** (0.055)
RT_10	0.838*** (0.058)
RT_11	0.731*** (0.059)
RT_12	0.756*** (0.059)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred. Pos
Obs. Neg	92.760	23.738
Obs. Pos	22.670	80.781

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	194.822	53.970
Obs. Pos	54.315	169.745

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.088^{***}
	(0.004)
Woe_SBIF_NINSDD_UM	0.108^{***}
	(0.006)
Woe_SBIF_DDIRVG_UM	0.172^{***}
	(0.007)
Woe_SBIF_MNCRDI_UM	0.540^{***}
	(0.014)
Woe_MesAntiguedadCliente	0.265^{***}
	(0.020)
Woe_EdadPeriodo	0.009
	(0.077)
Woe_SEXO	0.924^{***}
	(0.056)
Woe_region_zona	0.588^{***}
	(0.053)
Woe_RECENCIA_RETAIL	0.078^{***}
	(0.009)
Woe_RECENCIA_TARJETA	0.395^{***}
	(0.010)
Woe_porc_uso_adj	0.248^{***}
	(0.005)
Woe_Deuda	0.466^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	0.131^{***}
	(0.005)
Woe_mark_trx_AE	0.277^{***}
	(0.017)
Woe_mark_trx_SAE	0.285^{***}
	(0.036)
Woe_mark_trx_seguros_TCC	0.747^{***}
	(0.017)
Woe_mark_trx_seguros_TCA	0.504^{***}
	(0.011)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target 9 meses
Matrices de confusión target 9 meses.

Cabe destacar, que los datos totales en el testeo disminuyeron debido a que a medida que pasa el tiempo, empiezan a faltar meses disponibles para predecir. Entonces, los datos de test de predicción para el 9no mes llegan hasta abril de 2019.

Tabla 27. Matriz de confusión datos de testeo, target 9 meses.

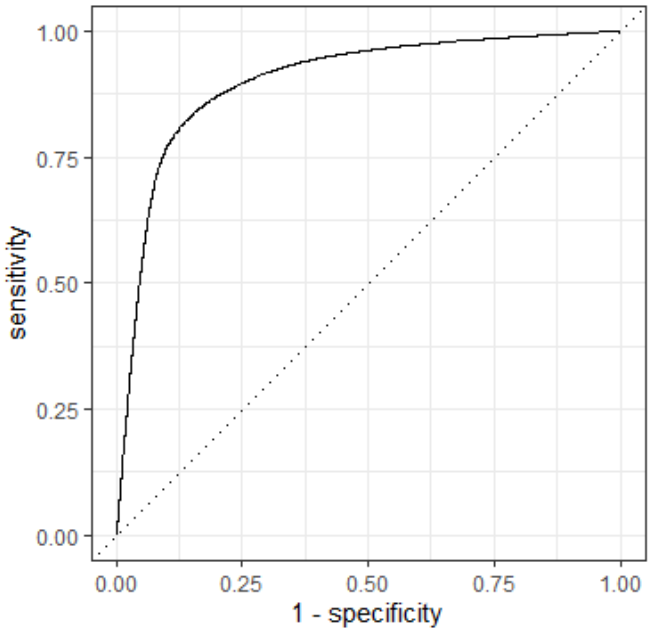
	Reales		
Predichos		0	1
	0	74.558	17.589
	1	18.990	64.725

Tabla 28. Matriz de confusión datos de entrenamiento, target 9 meses.

	Reales		
Predichos		0	1
	0	202.436	52.330
	1	47.831	169.867

Se procede a entregar las curvas ROC junto a su área bajo la curva del target a 9 meses.

Gráfico 31: Curva ROC para target a 9 meses, datos de testeo.

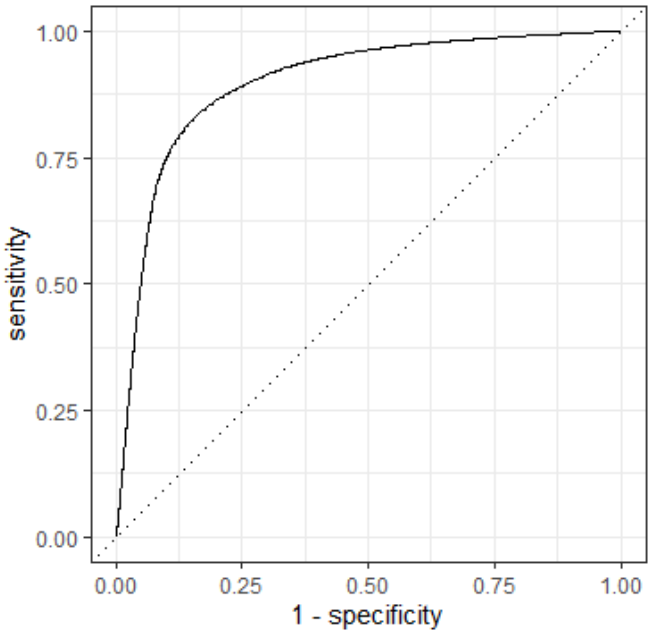


Fuente: Elaboración propia, a través de paquete parsnip en R.

El área bajo la curva de esta ROC es de 0,823 que indica que aun cuando ha disminuido este valor respecto a los anteriores, el modelo funciona bien con los datos que se le está entregando.

Para compararlo con el training set:

Gráfico 32: Curva ROC para target a 9 meses, datos de entrenamiento.



Fuente: Elaboración propia, a través de paquete parsnip en R.

El área bajo la curva es de 0,823. Se puede decir entonces que considerando el tiempo al que está pronosticando, el modelo homogéneo para esta variable independiente sirve para predecir el periodo t+9.

Tabla 29. Tabla de coeficientes regresión logística, target 9 meses.

	Variable Dependiente:
	Ingresos
(Intercept)	-5.133^{***}
	(0.081)
DDIRVG_norm	0.701
	(0.574)
NINSDD_norm	2.957^{***}
	(0.042)
DDAHIP_norm	-4.220^{***}
	(0.558)
MNCRDI_norm	-3.909^{***}
	(0.383)
EdadPeriodo_norm	0.511^{***}
	(0.026)
MesAntiguedadCliente_norm	0.209^{***}
	(0.037)
Porc_uso_norm	1.460^{***}
	(0.020)
Deuda_norm	10.735^{***}
	(0.335)
mark_trx_seguros_TCC	2.642^{***}
	(0.044)
mark_trx_seguros_TCA	1.959^{***}
	(0.035)
mark_trx_SAE	0.672^{**}
	(0.136)
mark_trx_AE	0.487^{***}
	(0.057)
Enero	-0.052^{**}
	(0.020)
Febrero	0.030
	(0.020)
Abril	0.085^{***}
	(0.019)
Mayo	0.168^{***}
	(0.019)
Junio	0.096^{***}
	(0.020)
Julio	0.012
	(0.020)

Agosto	−0.013
	(0.020)
Septiembre	0.026
	(0.020)
Octubre	0.074***
	(0.019)
Noviembre	−0.005
	(0.020)
Forma_Interes	1.703***
	(0.044)
Forma_Mixto	1.897***
	(0.050)
Forma_Contado	1.229***
	(0.048)
Sexo_M	−0.127***
	(0.008)
Region_1	−0.014
	(0.038)
Region_2	0.016
	(0.026)
Region_3	0.086*
	(0.036)
Region_4	0.081**
	(0.025)
Region_5	−0.059**
	(0.018)
Region_6	−0.033
	(0.024)
Region_7	−0.086***
	(0.023)
Region_8	−0.000
	(0.018)
Region_9	−0.042.
	(0.024)
Region_10	0.009
	(0.026)
Region_11	0.037
	(0.112)
Region_12	0.069
	(0.060)
Region_14	−0.112*
	(0.048)
Region_15	0.165***
	(0.036)

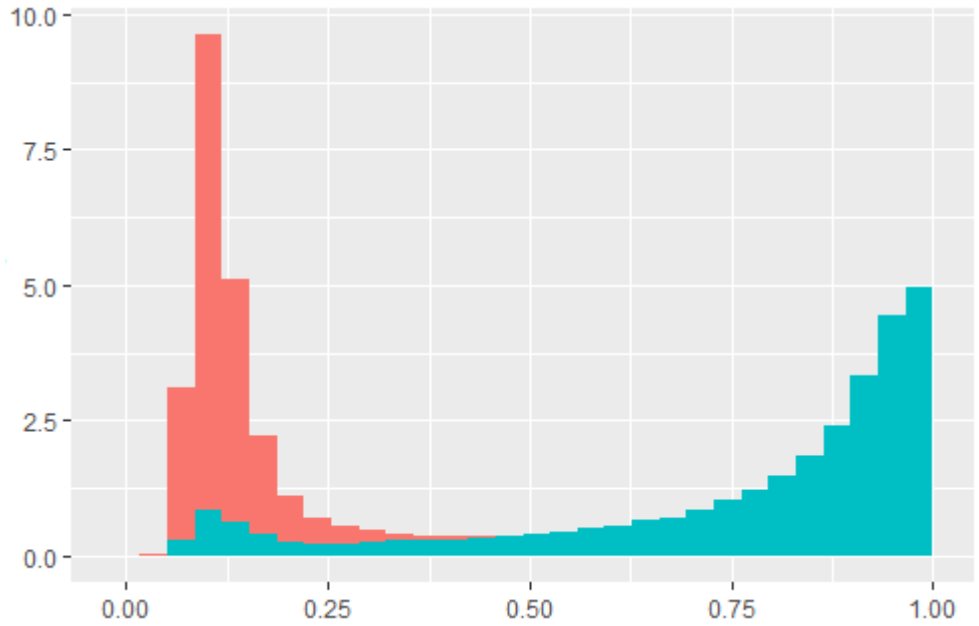
RM_CentroNorte	-0.064^{***}
	(0.019)
RM_Occidente	0.020
	(0.020)
RM_Oriente	-0.150^{***}
	(0.021)
RM_Periferica	0.061.
	(0.034)
RM_Sur	0.038^{*}
	(0.019)
RR_1	0.321^{***}
	(0.038)
RR_2	0.410^{***}
	(0.040)
RR_3	0.317^{***}
	(0.044)
RR_4	0.382^{***}
	(0.045)
RR_5	0.423^{***}
	(0.047)
RR_6	0.237^{***}
	(0.049)
RR_7	0.232^{***}
	(0.050)
RR_8	0.178^{***}
	(0.053)
RR_9	0.184^{***}
	(0.054)
RR_10	0.185^{***}
	(0.055)
RR_11	0.135^{*}
	(0.057)
RR_12	0.024
	(0.057)
RT_1	1.529^{***}
	(0.059)
RT_2	0.757^{***}
	(0.061)
RT_3	0.569^{***}
	(0.063)
RT_4	1.635^{***}
	(0.045)

RT_5	1.414*** (0.048)
RT_6	1.366*** (0.049)
RT_7	1.219*** (0.050)
RT_8	1.134*** (0.054)
RT_9	1.009*** (0.055)
RT_10	0.836*** (0.057)
RT_11	0.772*** (0.058)
RT_12	0.805*** (0.058)
Note:	*p<0.1; **p<0.05; ***p<0.01

En este último se mantienen las mismas características que el modelo a un target de 4 meses. Cabe destacar que como cada modelo es independiente, es normal que se encuentren distintas conclusiones por modelo. (en la entrega final, se plantea entregar una tabla con todas las variables y cómo se comportan)

Al igual que en los modelos homogéneos anteriores, se entregan los resultados del cómo está separando el modelo las probabilidades:

Gráfico 33: Separación del modelo a target de 9 meses.



Fuente: Elaboración propia.

Al igual que al target de 4 meses, al modelo le cuesta muy poco separar las variables, por lo que aun con el paso del tiempo logra separar las variables sin mayores dificultades.

A continuación, están los resultados obtenidos por el WoE:

Matrices de confusión WoE

Tabla 30. Matriz de confusión datos de testeo, target 9 meses WoE.

	Reales		
Predichos		0	1
	0	74.466	18.473
	1	19.224	63.997

Tabla 31. Matriz de confusión datos de entrenamiento, target 9 meses WoE.

	Reales		
Predichos		0	1
	0	195.025	55.608
	1	56.014	166.322

A continuación, la tabla de los coeficientes de la regresión con WoE:

Tabla 32. Tabla de coeficientes regresión logística WoE, target 9 meses.

	Variable Dependiente:
	Ingresos
(Intercept)	-0.208^{***}
	(0.008)
Woe_SBIF_NINSDD_UM	0.068^{***}
	(0.004)
Woe_SBIF_DDIRVG_UM	-0.206^{***}
	(0.006)
Woe_SBIF_MNCRDI_UM	-0.539^{***}
	(0.014)
Woe_MesAntiguedadCliente	-0.261^{***}
	(0.020)
Woe_EdadPeriodo	0.132.
	(0.078)
Woe_SEXO	-0.913^{***}
	(0.053)
Woe_region_zona	-0.680^{***}
	(0.049)
Woe_RECENCIA_RETAIL	-0.091^{***}
	(0.009)
Woe_RECENCIA_TARJETA	-0.378^{***}
	(0.010)
Woe_porc_uso_adj	-0.215^{***}
	(0.005)
Woe_Deuda	-0.474^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	-0.139^{***}
	(0.005)
Woe_mes	-1.800^{***}
	(0.154)
Woe_mark_trx_AE	-0.278^{***}
	(0.016)
Woe_mark_trx_SAE	-0.351^{***}
	(0.037)
Woe_mark_trx_seguros_TCC	-0.740^{***}
	(0.016)
Woe_mark_trx_seguros_TCA	-0.478^{***}
	(0.011)
Note:	*p<0.1; **p<0.05; ***p<0.01

A diferencia del primer modelo, se ve que aquí baja la significancia de la edad que se tiene en el periodo y que la toma de seguros o avances en efectivo tiene un efecto negativo a la hora de volver a comprar.

Target: 10 meses.

Matrices de confusión
Matriz de confusión de prueba

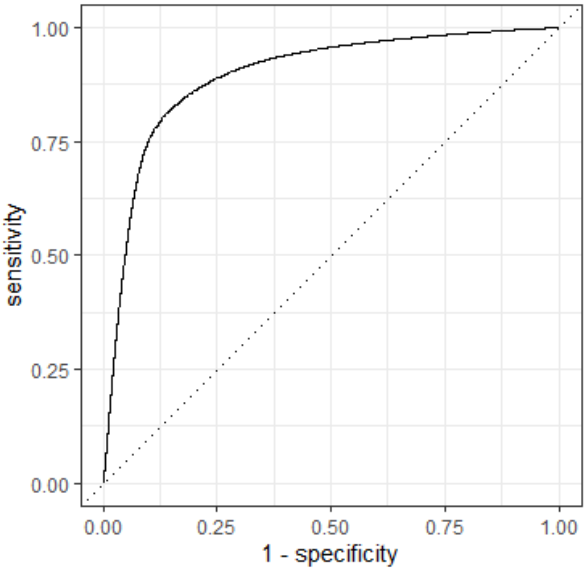
		Truth	
		0	1
Prediction	0	50.534	18.372
	1	11.332	50.619

Matriz de confusión train set

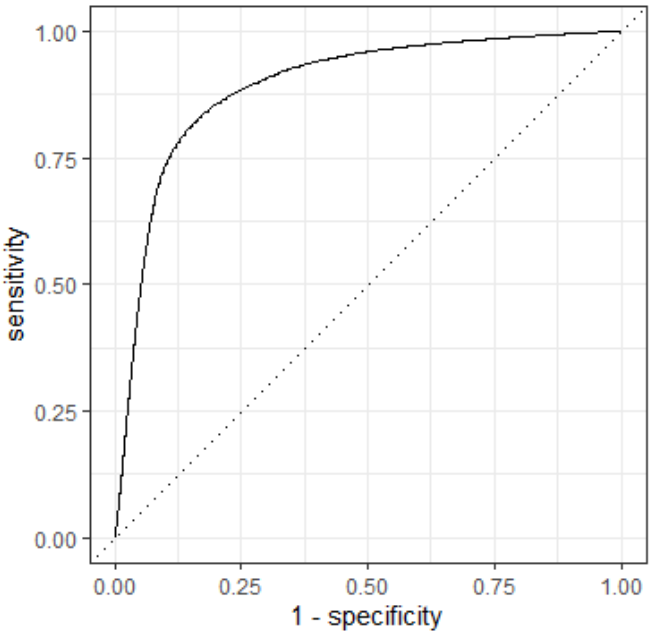
		Truth	
		0	1
Prediction	0	191.734	64.980
	1	60.485	156.251

ROC y AUC

Test: AUC: 0,820



Train: AUC: 0,821.



Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-5.503^{***}
	(0.091)
DDIRVG_norm	-0.408
	(0.898)
NINSDD_norm	2.851^{***}
	(0.041)
DDAHIP_norm	-2.018^{***}
	(0.368)
MNCRDI_norm	-3.573^{***}
	(0.355)
EdadPeriodo_norm	0.491^{***}
	(0.026)
MesAntiguedadCliente_norm	0.202^{***}
	(0.035)
Porc_uso_norm	1.293^{***}
	(0.019)
Deuda_norm	12.646^{***}
	(0.389)
mark_trx_seguros_TCC	2.460^{***}
	(0.042)
mark_trx_seguros_TCA	1.857^{***}
	(0.033)
mark_trx_SAE	0.325^{**}
	(0.113)
mark_trx_AE	0.357^{***}
	(0.052)
Enero	-0.021
	(0.019)
Febrero	0.050^{*}
	(0.019)
Abril	0.227^{***}
	(0.019)
Mayo	0.182^{***}
	(0.019)
Junio	0.075^{***}
	(0.019)
Julio	0.079^{***}
	(0.019)

Agosto	0.097^{***}
	(0.019)
Septiembre	0.072^{***}
	(0.019)
Octubre	0.143^{***}
	(0.019)
Noviembre	0.088^{***}
	(0.019)
Forma_Interes	1.749^{***}
	(0.042)
Forma_Mixto	1.975^{***}
	(0.048)
Forma_Contado	1.302^{***}
	(0.045)
Sexo_M	-0.125^{***}
	(0.008)
Region_1	0.006
	(0.037)
Region_2	0.003
	(0.026)
Region_3	0.076^{*}
	(0.035)
Region_4	0.044.
	(0.025)
Region_5	-0.073^{***}
	(0.018)
Region_6	-0.031
	(0.024)
Region_7	-0.108^{***}
	(0.023)
Region_8	-0.056^{**}
	(0.018)
Region_9	-0.052^{*}
	(0.024)
Region_10	0.070^{**}
	(0.026)
Region_11	0.179
	(0.110)
Region_12	-0.072
	(0.059)
Region_14	-0.119^{*}
	(0.047)
Region_15	0.091^{*}
	(0.035)

RM_CentroNorte	-0.044^{*}
	(0.018)
RM_Occidente	-0.014
	(0.019)
RM_Oriente	-0.157^{***}
	(0.021)
RM_Periferica	0.056.
	(0.034)
RM_Sur	0.006
	(0.019)
RR_1	0.306^{***}
	(0.037)
RR_2	0.266^{***}
	(0.040)
RR_3	0.251^{***}
	(0.043)
RR_4	0.332^{***}
	(0.044)
RR_5	0.323^{***}
	(0.046)
RR_6	0.334^{***}
	(0.048)
RR_7	0.215^{***}
	(0.049)
RR_8	0.191^{***}
	(0.052)
RR_9	0.074
	(0.053)
RR_10	0.117^{*}
	(0.055)
RR_11	0.031
	(0.056)
RR_12	0.094.
	(0.057)
RT_1	1.436^{***}
	(0.057)
RT_2	0.792^{***}
	(0.059)
RT_3	0.540^{***}
	(0.060)
RT_4	1.685^{***}
	(0.044)

RT_5	1.495*** (0.047)
RT_6	1.306*** (0.048)
RT_7	1.219*** (0.050)
RT_8	1.075*** (0.052)
RT_9	1.091*** (0.054)
RT_10	0.916*** (0.056)
RT_11	0.856*** (0.057)
RT_12	0.740*** (0.058)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	pred. Pos
Obs. Neg	55.618	14.537
Obs. Pos	13.941	47.745

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	197.150	56.247
Obs. Pos	55.589	164.898

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	-0.126^{***}
	(0.004)
Woe_SBIF_NINSDD_UM	0.110^{***}
	(0.006)
Woe_SBIF_DDIRVG_UM	0.190^{***}
	(0.007)
Woe_SBIF_MNCRDI_UM	0.518^{***}
	(0.013)
Woe_MesAntiguedadCliente	0.283^{***}
	(0.019)
Woe_EdadPeriodo	0.171^{*}
	(0.071)
Woe_SEXO	0.918^{***}
	(0.052)
Woe_region_zona	0.661^{***}
	(0.049)
Woe_RECENCIA_RETAIL	0.098^{***}
	(0.009)
Woe_RECENCIA_TARJETA	0.374^{***}
	(0.010)
Woe_porc_uso_adj	0.212^{***}
	(0.005)
Woe_Deuda	0.453^{***}
	(0.003)
Woe_MR_FORMA_TRX_U3M	0.145^{***}
	(0.005)
Woe_mark_trx_AE	0.263^{***}
	(0.016)
Woe_mark_trx_SAE	0.313^{***}
	(0.033)
Woe_mark_trx_seguros_TCC	0.752^{***}
	(0.016)
Woe_mark_trx_seguros_TCA	-0.5485^{***}
	(0.010)
Note:	*p<0.1; **p<0.05; ***p<0.01

Target: 11 meses.
Matrices de confusión

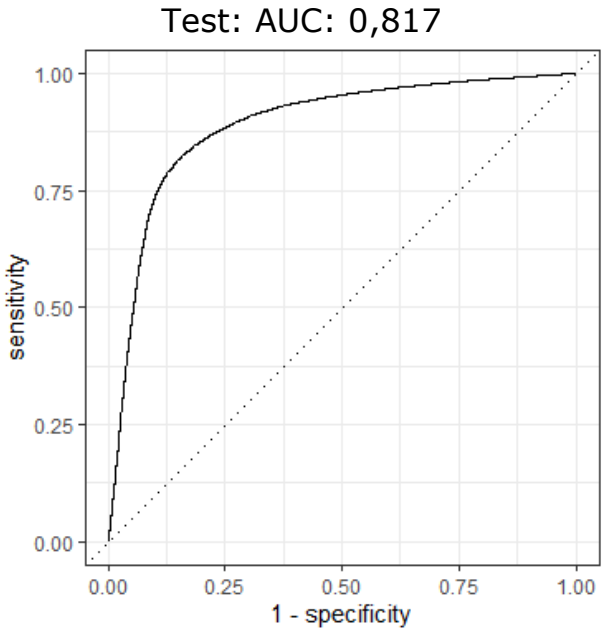
Matriz de confusión de prueba

		Truth	
		0	1
Prediction	0	29.560	12.311
	1	10.699	34.232

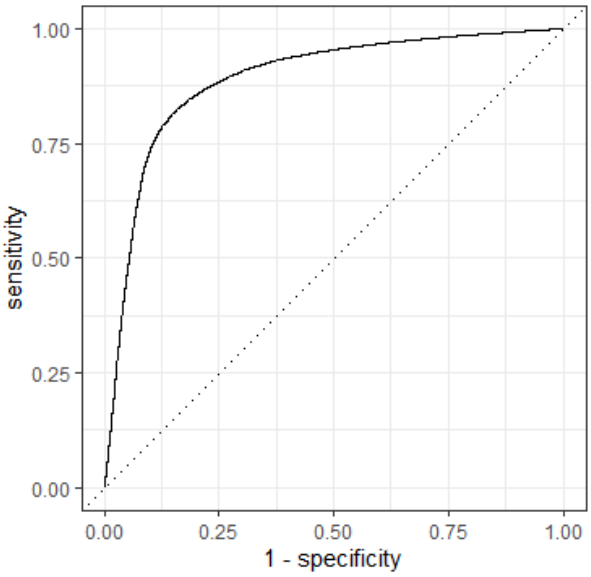
Matriz de confusión train set

		Truth	
		0	1
Prediction	0	149.928	79.380
	1	74.935	169.110

ROC y AUC



Train: AUC:0,817



Coeficientes

	Variable Dependiente:
	Ingresos
(Intercept)	-6.469^{***}
	(0.124)
DDIRVG_norm	0.957
	(2.660)
NINSDD_norm	2.633^{***}
	(0.040)
DDAHIP_norm	-2.569^{***}
	(0.263)
MNCRDI_norm	-3.175^{***}
	(0.355)
EdadPeriodo_norm	0.386^{***}
	(0.025)
MesAntiguedadCliente_norm	0.341^{***}
	(0.037)
Porc_uso_norm	1.031^{***}
	(0.018)
Deuda_norm	10.809^{***}
	(0.336)
mark_trx_seguros_TCC	2.155^{***}
	(0.038)
mark_trx_seguros_TCA	1.621^{***}
	(0.030)
mark_trx_SAE	0.347^{***}
	(0.100)
mark_trx_AE	0.310^{***}
	(0.045)
Enero	0.001
	(0.019)
Febrero	0.143^{***}
	(0.019)
Abril	0.088^{***}
	(0.019)
Mayo	0.068^{***}
	(0.019)
Junio	0.047^{*}
	(0.019)
Julio	0.053^{**}
	(0.019)

Agosto	0.042^{*}
	(0.019)
Septiembre	0.086^{***}
	(0.019)
Octubre	0.091^{***}
	(0.019)
Noviembre	0.003
	(0.019)
Forma_Interes	1.639^{***}
	(0.038)
Forma_Mixto	1.874^{***}
	(0.044)
Forma_Contado	1.275^{***}
	(0.042)
Sexo_M	-0.117^{***}
	(0.008)
Region_1	-0.072^{*}
	(0.036)
Region_2	0.044.
	(0.025)
Region_3	0.120^{***}
	(0.034)
Region_4	0.051^{*}
	(0.024)
Region_5	-0.037^{*}
	(0.017)
Region_6	0.010
	(0.023)
Region_7	-0.112^{***}
	(0.022)
Region_8	-0.021
	(0.017)
Region_9	-0.018
	(0.023)
Region_10	0.059^{*}
	(0.025)
Region_11	-0.067
	(0.108)
Region_12	-0.056
	(0.058)
Region_14	-0.168^{***}
	(0.046)
Region_15	0.138^{***}
	(0.035)

RM_CentroNorte	-0.051^{**}
	(0.018)
RM_Occidente	0.018
	(0.019)
RM_Oriente	-0.153^{***}
	(0.020)
RM_Periferica	0.111^{***}
	(0.033)
RM_Sur	0.000
	(0.018)
RR_1	0.432^{***}
	(0.034)
RR_2	0.354^{***}
	(0.037)
RR_3	0.272^{***}
	(0.040)
RR_4	0.305^{***}
	(0.041)
RR_5	0.327^{***}
	(0.043)
RR_6	0.223^{***}
	(0.045)
RR_7	0.231^{***}
	(0.047)
RR_8	0.175^{***}
	(0.049)
RR_9	0.205^{***}
	(0.051)
RR_10	0.085
	(0.053)
RR_11	0.055
	(0.054)
RR_12	0.013
	(0.055)
RT_1	1.273^{***}
	(0.053)
RT_2	0.713^{***}
	(0.054)
RT_3	0.488^{***}
	(0.056)
RT_4	1.614^{***}
	(0.042)

RT_5	1.442*** (0.044)
RT_6	1.276*** (0.046)
RT_7	1.138*** (0.048)
RT_8	1.015*** (0.050)
RT_9	0.922*** (0.052)
RT_10	0.824*** (0.055)
RT_11	0.798*** (0.055)
RT_12	0.729*** (0.056)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión WoE

Test:

	Pred. Neg	Pred.pos
Obs. Neg	29.202	12.054
Obs. Pos	11.645	35.159

Train:

	Pred. Neg	Pred. Pos
Obs. Neg	198.807	55.173
Obs. Pos	70.339	148.452

Coeficientes WoE

	Variable Dependiente:
	Ingresos
(Intercept)	−0.160***
	(0.004)
Woe_SBIF_NINSDD_UM	−0.111***
	(0.006)
Woe_SBIF_DDIRVG_UM	−0.185***
	(0.007)
Woe_SBIF_MNCRDI_UM	−0.528***
	(0.012)
Woe_MesAntiguedadCliente	−0.282***
	(0.019)
Woe_EdadPeriodo	−0.249***
	(0.065)
Woe_SEXO	−0.907***
	(0.046)
Woe_region_zona	−0.680***
	(0.048)
Woe_RECENCIA_RETAIL	−0.097***
	(0.009)
Woe_RECENCIA_TARJETA	−0.369***
	(0.009)
Woe_porc_uso_adj	−0.182***
	(0.005)
Woe_Deuda	−0.446***
	(0.003)
Woe_MR_FORMA_TRX_U3M	−0.158***
	(0.005)
Woe_mark_trx_AE	−0.253***
	(0.015)
Woe_mark_trx_SAE	−0.304***
	(0.032)
Woe_mark_trx_seguros_TCC	−0.696***
	(0.016)
Woe_mark_trx_seguros_TCA	−0.467***
	(0.010)
Note:	*p<0.1; **p<0.05; ***p<0.01

Matrices de confusión de modelos benchmark

Matriz de confusión primer modelo unitario, test set.

	Reales		
Predichos		0	1
	0	12.509	7.904
	1	8.074	34.376

Dando entonces un accuracy de 72,2%. Ahora con el entrenamiento:

Matriz de confusión datos de entrenamiento, segundo modelo unitario.

	Reales		
Predichos		0	1
	0	19.500	13.849
	1	11.214	43.724

Con un accuracy total de 71,4%. Como se puede ver, la variable frecuencia logra explicar de buena manera la compra o no de clientes. Hay que considerar que estas variables están correlacionadas en la lógica simple.

Matriz de confusión segundo modelo unitario, test set.

	Reales		
Predichos		0	1
	0	19.585	6.404
	1	11.456	27.499

Dando entonces un accuracy de 72,5%. Ahora con el entrenamiento:

Matriz de confusión datos de entrenamiento, segundo modelo unitario.

	Reales		
Predichos		0	1
	0	27.675	7.384
	1	6.129	62.574

Con un accuracy total de 71,7%.

Matriz de confusión tercer modelo unitario, test set.

	Reales		
Predichos		0	1
	0	21.177	9.178
	1	7.448	27.141

Con un resultado de 74,4%, tiene una leve mejora con respecto al primer intento. Ahora con el entrenamiento:

Matriz de confusión datos de entrenamiento, tercer modelo unitario.

	Reales		
Predichos		0	1
	0	29.102	11.707
	1	10.451	51.762

Entregando un resultado de 77,9% de accuracy.

Matriz de correlación con variable ingreso.

	Periodo	ntajeusoTa	Deuda	F_Deudavig	meroInsti	DeudaHipot	LineaDeCr	Edad	Ingreso	iguedadCli	Frecuencia
Periodo	1.0000	-0.0088	0.0311	0.0200	0.1991	0.0196	0.0447	-0.0064	0.0095	-0.0083	-0.0762
PorcentajeUsoTarjeta	-0.0088	1.0000	0.4517	-0.0100	0.2907	-0.0887	-0.1972	-0.0846	0.3141	-0.1043	0.3804
Deuda	0.0311	0.4517	1.0000	0.1046	0.2824	0.0191	0.0179	0.0199	0.5367	0.0796	0.2745
Sbif_Deudavigente	0.0200	-0.0100	0.1046	1.0000	0.2201	0.6364	0.3523	-0.0922	0.0590	0.0372	0.0159
Sbif_NumeroInstitucion	0.1991	0.2907	0.2824	0.2201	1.0000	0.1720	0.2315	-0.1381	0.1829	0.0614	0.3284
Sbif_DeudaHipotecaria	0.0196	-0.0887	0.0191	0.6364	0.1720	1.0000	0.4200	-0.1628	0.0032	0.0185	-0.0370
Sbif_LineaDeCredito	0.0447	-0.1972	0.0179	0.3523	0.2315	0.4200	1.0000	0.0021	-0.0187	0.1600	0.0089
Edad	-0.0064	-0.0846	0.0199	-0.0922	-0.1381	-0.1628	0.0021	1.0000	-0.0054	0.3697	0.1051
Ingreso	0.0095	0.3141	0.5367	0.0590	0.1829	0.0032	-0.0187	-0.0054	1.0000	0.0094	0.1776
AntiguedadCliente	-0.0083	-0.1043	0.0796	0.0372	0.0614	0.0185	0.1600	0.3697	0.0094	1.0000	0.2653
Frecuencia	-0.0762	0.3804	0.2745	0.0159	0.3284	-0.0370	0.0089	0.1051	0.1776	0.2653	1.0000

Matriz de correlación variable margen

	Periodo	Margen	Sbif_Deuda	Sbif_Numer	Sbif_Deuda	Sbif_Linea	Edad	PorcentajeDeuda	Antiguedad	Frencuenci
Periodo	1.0000	-0.0061	0.0174	0.1843	0.0129	0.0410	0.0395	-0.0238	0.0710	0.0002
Margen	-0.0061	1.0000	0.0102	0.0333	-0.0006	-0.0045	0.0086	0.0651	0.0135	0.0812
Sbif_DeudaVigente	0.0174	0.0102	1.0000	0.2483	0.7601	0.4474	-0.1407	-0.0061	0.0186	-0.0196
Sbif_NumeroInstituci	0.1843	0.0333	0.2483	1.0000	0.1753	0.2176	-0.1647	0.3370	0.0597	0.3799
Sbif_DeudaHipotecaria	0.0129	-0.0006	0.7601	0.1753	1.0000	0.4193	-0.1772	-0.0642	0.0065	-0.0569
Sbif_LineaDeCredito	0.0410	-0.0045	0.4474	0.2176	0.4193	1.0000	-0.0190	-0.1671	0.1194	-0.0862
Edad	0.0395	0.0086	-0.1407	-0.1647	-0.1772	-0.0190	1.0000	-0.0676	0.3218	0.0459
PorcentajeUsoTarjeta	-0.0238	0.0651	-0.0061	0.3370	-0.0642	-0.1671	-0.0676	1.0000	-0.0886	0.5365
Deuda	0.0215	0.0930	0.1135	0.2830	0.0229	0.0200	0.0150	0.4850	1.0000	0.2922
AntiguedadCliente	0.0710	0.0135	0.0186	0.0597	0.0065	0.1194	0.3218	-0.0886	0.0584	1.0000
Frencuencia	0.0002	0.0812	-0.0196	0.3799	-0.0569	-0.0862	0.0459	0.5365	0.2922	0.0954