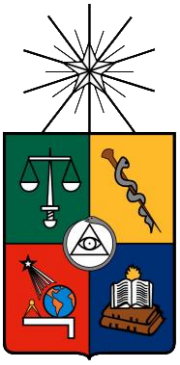


UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE RECOMENDACIÓN DE CLIENTES PARA LA GESTIÓN DE CAMPAÑAS DE PRODUCTOS DE CONSUMO PRE APROBADOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CAMILO ANDRÉS SALAZAR ORTEGA

PROFESOR GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:
ERICK MÉNDEZ GUZMÁN
CAROLINA SEGOVIA RIQUELME

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Camilo Andrés Salazar Ortega
FECHA: 09/04/2018
PROFESOR GUÍA: Alejandra Puente Chandía

MODELO DE OPTIMIZACIÓN DE LA GESTIÓN DE CAMPAÑAS DE PRODUCTOS DE CONSUMO PRE APROBADOS

El trabajo de título se lleva a cabo en un Banco privado que posee alrededor del 3% de las colocaciones totales en el sistema financiero (SBIF, 2017). El Banco actualmente se encuentra en un estado de crecimiento y uno de sus focos para los próximos años corresponde en aumentar las colocaciones de consumo.

El objetivo general del trabajo es desarrollar una metodología para aumentar la tasa de respuesta de las campañas de promoción de créditos de consumo y tarjetas de créditos pre aprobadas, para así aumentar las colocaciones de consumo.

La motivación de este trabajo radica en la oportunidad de optimizar la gestión de los ejecutivos en la promoción de productos pre aprobados, en este contexto, aproximadamente un 72% de los clientes es gestionado por los ejecutivos y teniendo en cuenta que la tasa de respuesta en créditos de consumo promedio de los clientes gestionados es 5,8% y 2,1% de los clientes que no son gestionados, es fundamental hacer una selección inteligente de los clientes a los que se les asigna una prioridad alta de gestión.

Para llevar a cabo el objetivo general se sigue la metodología CRISP-DM, donde se desarrolla un modelo de propensión de créditos de consumo, un modelo de propensión de tarjetas de crédito, una segmentación de clientes de acuerdo con el uso de correo electrónico y sitio web, se resuelve un problema de optimización lineal para realizar la priorización que deben seguir los ejecutivos y se plantea el diseño experimental para validar el aumento de la tasa de respuesta. Los modelos de propensión que se estudian son los árboles de decisión y regresiones logísticas en conjunto con las técnicas de balanceo de datos como Random Over Samplig, SMOTE y eliminación de enlaces de Tomek. En los modelos de segmentación se utiliza el algoritmo K-Means. Los modelos de propensión construidos permiten mejorar el sistema actual de priorización de clientes, alcanzando una mejora del 4,4% y 3,2% en la precisión de la priorización de créditos de consumo y tarjetas de crédito respectivamente. Y los modelos de segmentación permiten identificar a los clientes que se necesita contactar a través del ejecutivo, ya que estos no utilizan el correo electrónico o sitio web.

Al integrar los modelos de propensión y segmentación de uso de canales en un problema de optimización lineal se logra desarrollar una metodología que prioriza a los clientes que son más propensos a tomar los productos ofertados y que entregan una mayor utilidad esperada.

AGRADECIMIENTOS

Me gustaría agradecer a mi familia por el apoyo incondicional durante toda mi vida, por enseñarme una cantidad infinita de valores, por ayudarme a ver la vida de la mejor manera posible, por enseñarme a siempre dar lo máximo y por enseñarme el verdadero significado de la palabra familia.

A mi polola, Fernanda Poblete por el constante apoyo, confianza y paciencia en este largo y duro proceso, ya que en los momentos difíciles siempre ha estado presente.

Agradecer a mis amigos, los grandes amigos que conozco de toda la vida, José Miguel, Carolina, Álvaro, Rosario, Jaime, Francisca, Catalina, Raúl, Alonso, Sebastián, Michelle, Javiera, Gonzalo, Paulo y tantos otros con los que he vivido momentos inolvidables.

También agradecer a mis compañeros de universidad, Diego B., Felipe, Franco, Martin, Samu y muchos otros, con los que estudié, fui al gimnasio, jugué y seguiremos compartiendo.

También quiero agradecer a los profesores de la sección de Marketing Cuantitativo por su disposición y consejos durante el semestre, en especial a Alejandra Puente, por atender a mis consultas.

Finalmente, agradecer al área de inteligencia de negocios donde desarrolle el trabajo de memoria por la ayuda y la oportunidad de trabajar con ustedes.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1 Antecedentes generales

El tema de memoria es desarrollado en una empresa financiera que ofrece distintos servicios orientados a la banca personas, banca corporativa, mercado de capitales, administración de fondos, corredores de bolsa, inversiones y corredora de seguros, sin embargo, el tema se enfoca en la banca personas, ya que tiene relación con las campañas de promoción de productos de consumo pre aprobados, estos pueden ser créditos de consumo en cuotas, tarjetas de crédito y líneas de sobregiro.

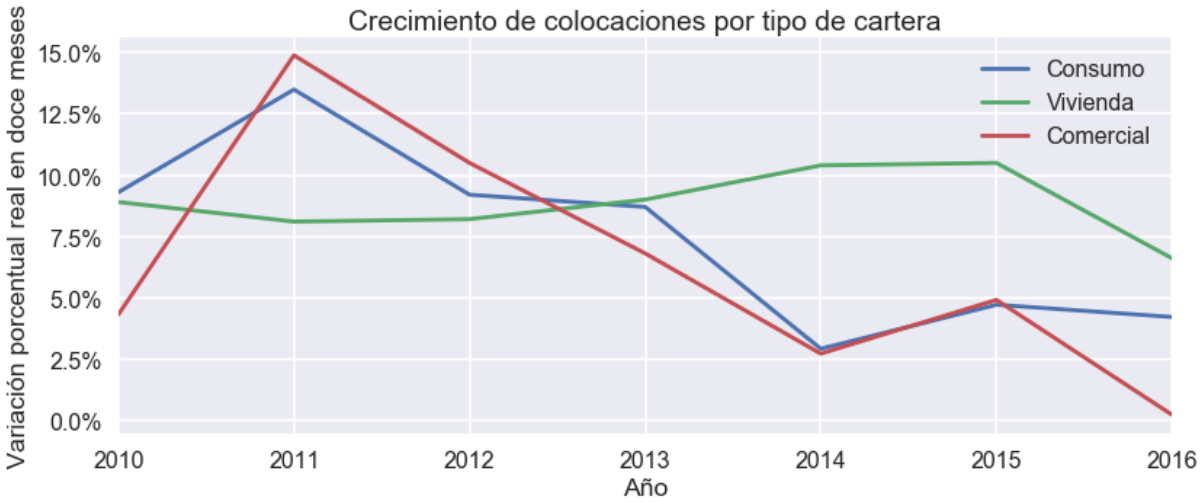
1.1 Industria bancaria

Los servicios ofrecidos por los bancos en general constan de depósitos, transacciones, préstamos, asesoramiento financiero, entre otros.

El año 2016 la industria bancaria posee 4,2 millones de clientes consumo, 1 millón de clientes vivienda, 1,3 millones de clientes pymes y 18 mil clientes de grandes empresas, generando un total de US$ 224 mil millones en colocaciones totales, lo que corresponde a un 85% del PIB (ABIF, 2016).

En la Ilustración 1, se tienen las variaciones históricas de las colocaciones por tipo de cartera de los últimos siete años. El año 2011 se tiene una variación porcentual real en doce meses del 13,5% en las colocaciones de consumo y del 14,9% en las colocaciones comerciales, posteriormente el año 2016 el crecimiento disminuye a un 4,2% y 0,2% respectivamente, por lo tanto, se observa que el crecimiento de estas carteras ha presentado una desaceleración muy violenta, a diferencia de la cartera hipotecaria que también ha sufrido una desaceleración, sin embargo, mucho menos pronunciada, ya que, entre los años 2011 y 2016 el crecimiento se mantuvo estable alrededor de una variación porcentual del 9% y el año 2016 llega a una variación porcentual del 6,6%.

Ilustración 1: Crecimiento de colocaciones por tipo de cartera en los últimos 7 años.

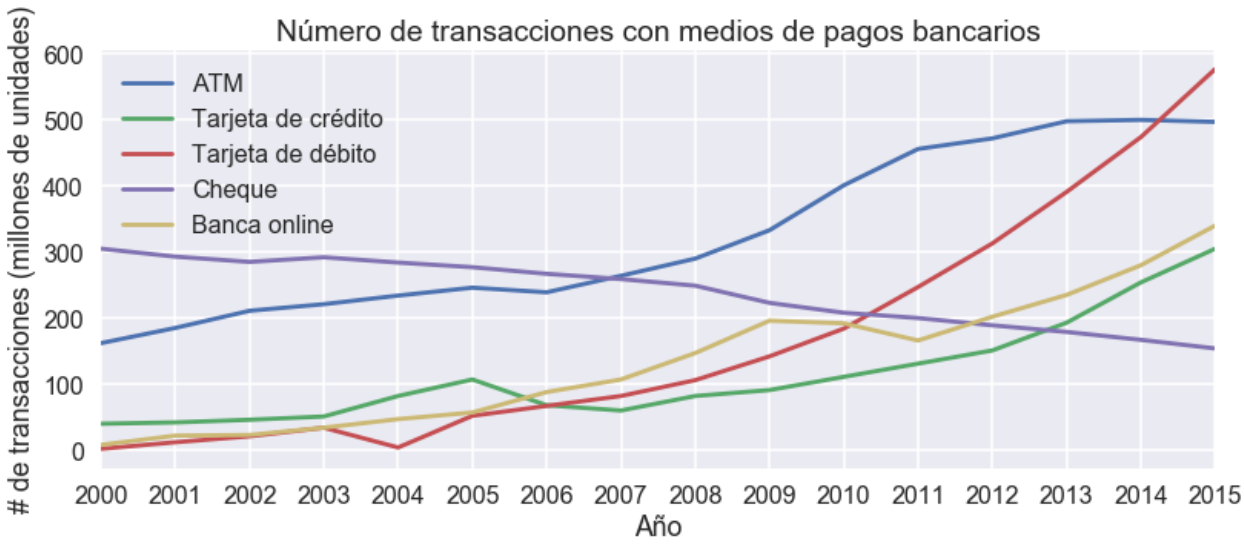


Fuente: Elaboración propia, datos ABIF (ABIF, 2016).

Con respecto al uso de los servicios financieros, se observa que el número de transacciones con medios de pagos bancarios en general ha ido en constante aumento y entre el año 2000 y el año 2015 se pasa de aproximadamente 517 millones de transacciones a 1.822 millones de transacciones, lo que se traduce en un aumento del 252% aproximadamente, este comportamiento se puede visualizar en la Ilustración 2.

Se debe destacar que en particular el número de transacciones en cajeros automáticos, tarjeta de crédito, tarjeta de débito y banca online ha aumentado, sin embargo, el uso de cheques ha ido en constante disminución.

Ilustración 2: Número de transacciones con diferentes medios de pagos bancarios en millones de unidades durante los últimos 16 años.



Fuente: Elaboración propia, datos ABIF (ABIF, 2016).

Al año 2017 existen 21 instituciones financieras activas, sin embargo, como se puede observar en la Tabla 1 entre las 5 entidades con mayor participación en las colocaciones totales del sistema financiero alcanzan el 78% de ellas, lo que revela que es una industria concentrada en pocos bancos.

Tabla 1: Porcentaje de participación en colocaciones totales por institución a marzo del 2017.

Institución	Marzo del 2017
Banco Santander-Chile	17,99%
Banco de Chile	16,85%
Banco de Crédito e Inversiones	14,82%
Banco del Estado de Chile	13,97%
Itaú Corpbanca (1)	13,95%
Banco Bilbao Vizcaya Argentaria, Chile	6,23%
Scotiabank Chile	6,15%
Banco Security	2,97%
Banco Bice	2,93%
Banco Consorcio	1,23%
Banco Fallabella	1,03%
Banco Internacional	0,63%
Banco Ripley	0,5%
Rabobank Chile	0,47%
HSBC Bank (Chile)	0,14%
Otros (4)	0,12%
Sistema	100,00%

Fuente: Elaboración propia con datos SBIF (SBIF, 2017).

En el ranking realizado por World Economic Forum (WEF, 2016) que incluye a 138 países, Chile se encuentra en la posición número 20 en cuanto a la facilidad al acceso de crédito bancario con una puntuación de 4,8 en una escala del 1 al 7, Chile se encuentra por sobre el promedio mundial de 3,9.

- Se puede concluir que Chile es un país con facilidad de acceso al crédito bancario, la industria financiera está concentrada en unos pocos bancos, el crecimiento de las colocaciones se ha visto desacelerado en los últimos años y en los últimos años el número de transacciones con todos los medios de pagos bancarios ha aumentado, a excepción del cheque.

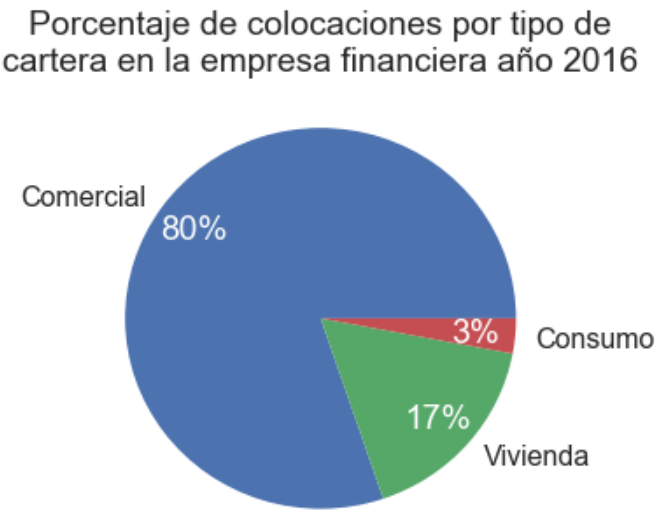
1.2 Banco

En el Banco en que se realiza la memoria, una parte importante de los ingresos proviene de las colocaciones. Estas corresponden a un préstamo de dinero que en el futuro el cliente devolverá con un interés adicional. Las colocaciones pueden ser clasificadas por tipo de deuda, esta puede ser de consumo, vivienda o comercial.

Según la memoria anual del año 2016, el Banco lidera en el crecimiento en el mercado hipotecario vivienda, con un aumento nominal en las colocaciones del 18%, sin embargo, no se obtienen los mismos resultados en la cartera de consumo y cartera comercial, donde sólo se ve un aumento del 4,4% y 5,5% respectivamente. En total, el crecimiento de las colocaciones del Banco el año 2016 es de un 7,4%.

Por otro lado, en la Ilustración 3 se aprecia que la mayor parte de las colocaciones del Banco se encuentran en la cartera comercial, con un 80,5% del total. Esto quiere decir que la mayor cantidad de sus colocaciones tiene relación con empresas.

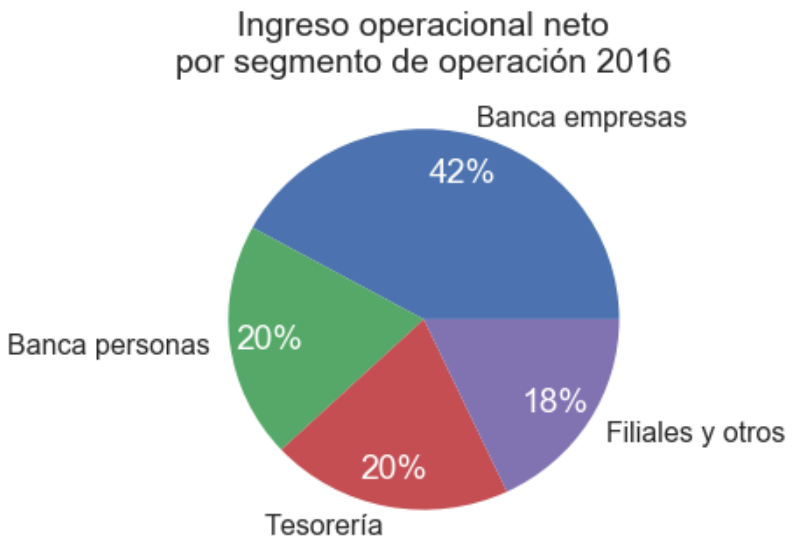
Ilustración 3: Porcentaje de colocaciones por tipo de cartera en la empresa financiera año 2016.



Fuente: Elaboración propia con datos de Memoria Anual de Banco 2016.

No obstante, al analizar los ingresos operacionales netos del año 2016 en la Ilustración 4, se puede observar que la distribución de ingresos no está concentrada exclusivamente en las empresas como si lo es en la distribución de las colocaciones, por lo que la cartera de consumo es una parte importante en los ingresos del Banco.

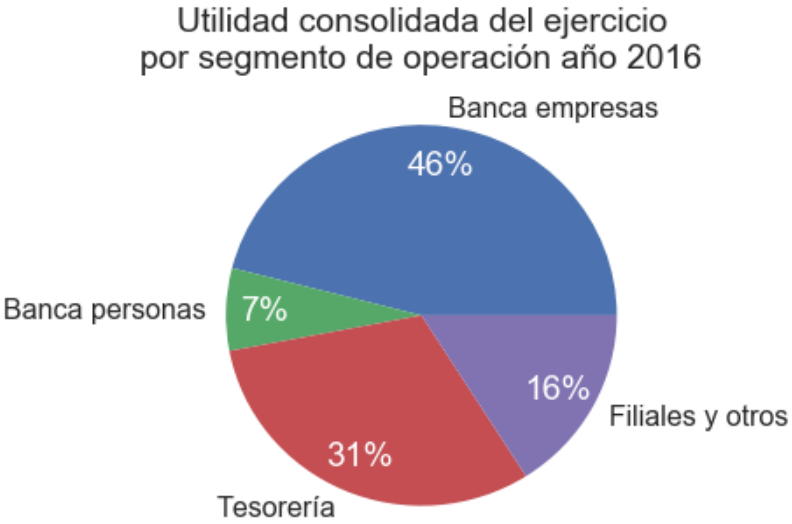
Ilustración 4: Distribución porcentual del ingreso operacional neto año 2016.



Fuente: Elaboración propia con datos de Memoria Anual de Banco 2016.

Finalmente, la utilidad consolidada del ejercicio de la banca personas decae luego de descontar los gastos operacionales y el impuesto a la renta, lo cual se observa en la Ilustración 5, ya que, sólo el 7% de las utilidades proviene de la banca personas.

Ilustración 5: Distribución porcentual de la utilidad consolidad del ejercicio año 2016.



Fuente: Elaboración propia con datos de Memoria Anual de Banco 2016.

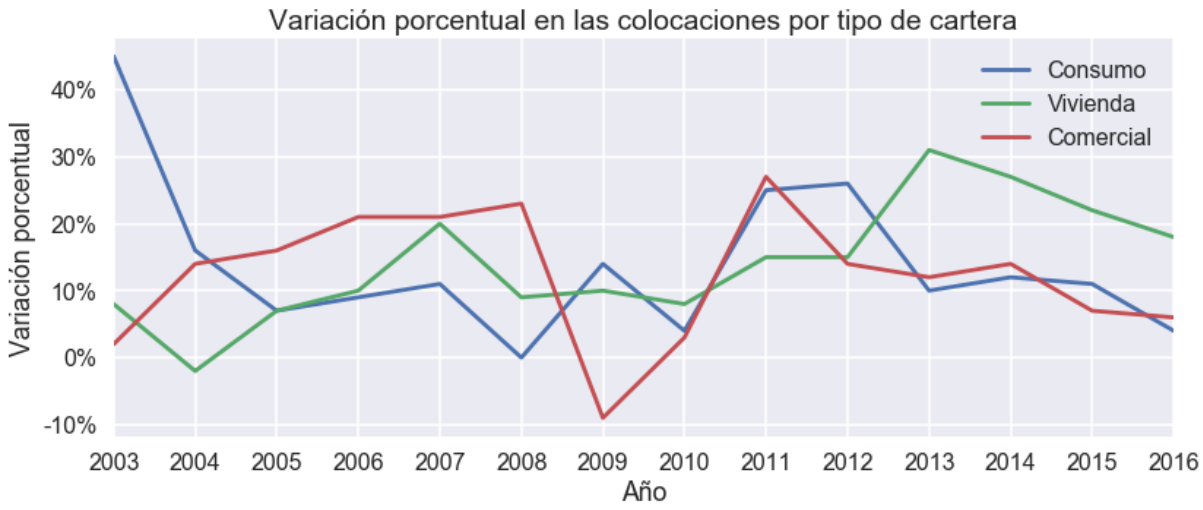
- Por lo tanto, en una industria que se encuentra en crecimiento, el Banco tiene el desafío de aumentar el crecimiento de las colocaciones de consumo y comerciales para así también liderar en el crecimiento de estas carteras. También, es importante aumentar las ventas de las colocaciones de consumo optimizando la utilización de los recursos para disminuir los gastos operacionales.

2 Descripción y justificación del proyecto

El problema que se identifica en el Banco tiene relación con la evolución de las colocaciones de la cartera de consumo de los últimos 5 años.

Se puede observar en la Ilustración 6, que la variación porcentual en el Banco de la cartera de consumo, hipotecaria y comercial ha sido positiva la mayoría de los años. Esto quiere decir que las colocaciones del Banco han permanecido en constante crecimiento, sin embargo, al analizar los últimos 5 años, el crecimiento de la cartera de consumo ha ido en constante desaceleración, pasando de un crecimiento del 25,7% el año 2012 a un 4,1% el año 2016, por lo tanto, se reduce en un 84% el crecimiento mientras que en la industria este porcentaje es de 53% (9,2% en año 2012 a 4,2% en año 2016, ver Ilustración 1).

Ilustración 6: Variación porcentual en las colocaciones por tipo de cartera del Banco.

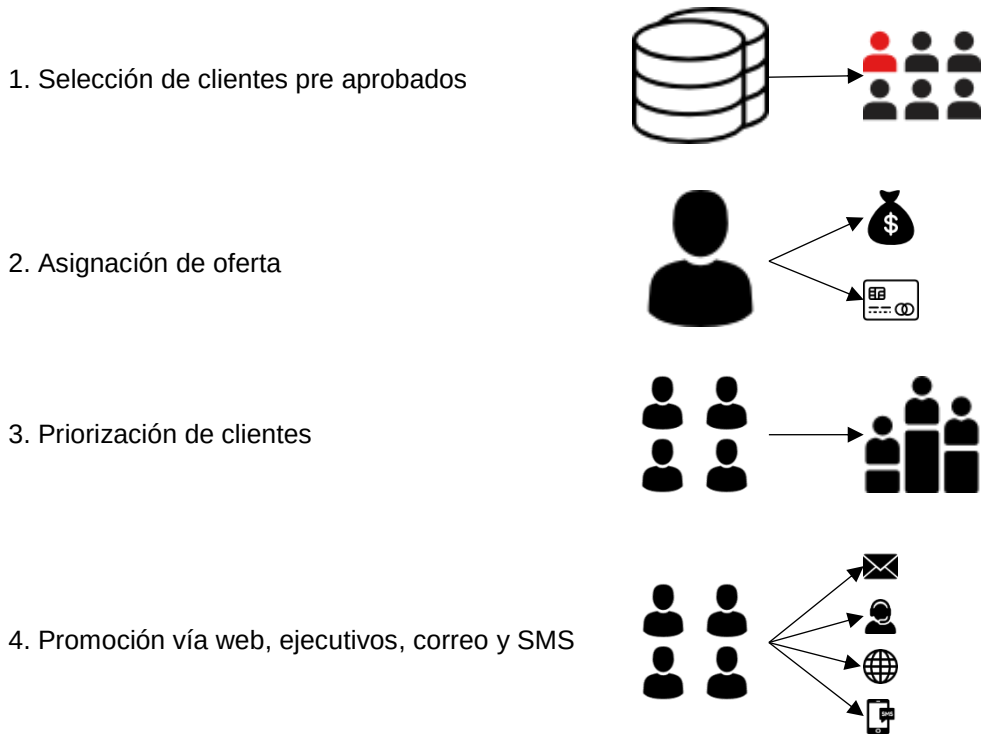


Fuente: Elaboración propia con recopilación de datos de Memorias Anuales del Banco.

Como se hizo notar en los antecedentes generales en la Ilustración 4 e Ilustración 5, aproximadamente el 20% de los ingresos proviene de la banca de personas, es por esta razón que el crecimiento de las colocaciones de consumo es muy relevante para el Banco.

2.1 Proceso de pre aprobación

Ilustración 7: Esquema del proceso de pre aprobación



Fuente: Elaboración propia.

La oferta de productos de consumo pre aprobados en el Banco se ha implementado desde el año 2015, estos corresponden a créditos de consumo, aumentos de cupo y apertura de tarjetas de crédito y líneas de sobregiro.

El proceso de pre aprobación se puede visualizar en la Ilustración 7, este comienza con la selección de clientes. Estos poseen una antigüedad mínima de seis meses, un comportamiento de pago excelente y una capacidad de endeudamiento mínima, por lo tanto, estos clientes pueden aumentar su endeudamiento a corto o largo plazo y el Banco está dispuesto a pre aprobar un producto de consumo con este tipo de clientes.

A cada cliente pre aprobado se le puede ofertar un crédito de consumo, línea de sobregiro y/o tarjeta de crédito, en el caso de la línea de sobregiro y tarjeta crédito la oferta puede ser un aumento de cupo o la apertura del producto.

Luego los clientes pueden realizar la activación automáticamente de los productos que tienen pre aprobados, cabe destacar que pueden activar un monto inferior al de la oferta recibida y transferir los montos entre los productos, por ejemplo, si el cliente no necesita una tarjeta de crédito el monto ofertado en la tarjeta puede traspasarse a crédito de consumo.

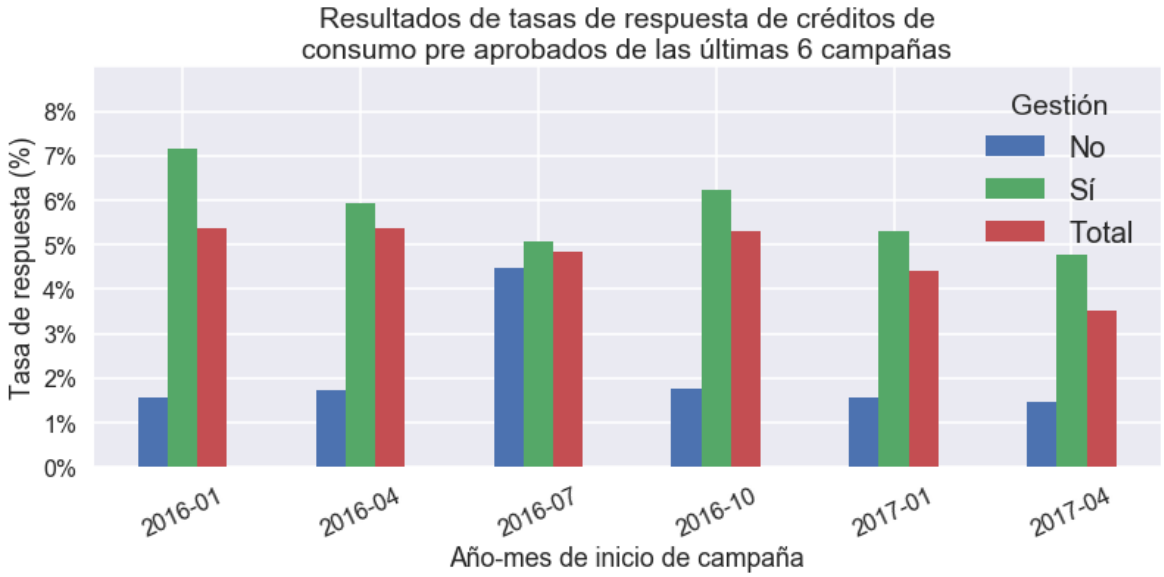
Los canales que se utilizan para hacer la promoción de esta oferta son: sitio web (los clientes al iniciar sesión en el sitio web privado reciben la oferta), e-mail, ejecutivos y en la campaña de agosto de 2017 se comienza a utilizar el envío de SMS a clientes que no usan el sitio web privado. No existe selección de canales exclusivos, todas las ofertas se realizan por sitio web, e-mail y ejecutivo, sin embargo, existen clientes que no abren los correos, una parte de los clientes dejó de suscribirse al contacto vía e-mail, no todos los clientes utilizan la página web, los mensajes de texto no siempre son recibidos o abiertos y los ejecutivos no son capaces de gestionar el 100% de los clientes. Por lo que es fundamental optimizar la utilización de los canales.

El Banco realiza una priorización de los clientes con oferta de acuerdo con reglas de tenencias de productos, segmento e información transaccional del cliente, por ejemplo, a clientes de un determinado segmento se les otorga una mayor prioridad, esto quiere decir que los ejecutivos se esforzarán en gestionar a estos clientes por sobre los demás.

La priorización se hace con el objetivo de aumentar la probabilidad de gestión y de venta, sin embargo, los resultados no son los esperados y se puede mejorar haciendo uso de la gran cantidad de información que el Banco posee acerca de estos clientes a través de la minería de datos.

Los resultados de las tasas de respuesta de créditos de consumo pre aprobados de las últimas 6 campañas se pueden observar en la Ilustración 8. La tasa de respuestas promedio de créditos de consumo desde enero de 2016 a abril de 2017 es de un 4,7%. También se puede apreciar que existe una diferencia significativa entre los clientes gestionados y los no gestionados, ya que, en promedio la tasa de respuesta es de 5,8% y 2,1% respectivamente.

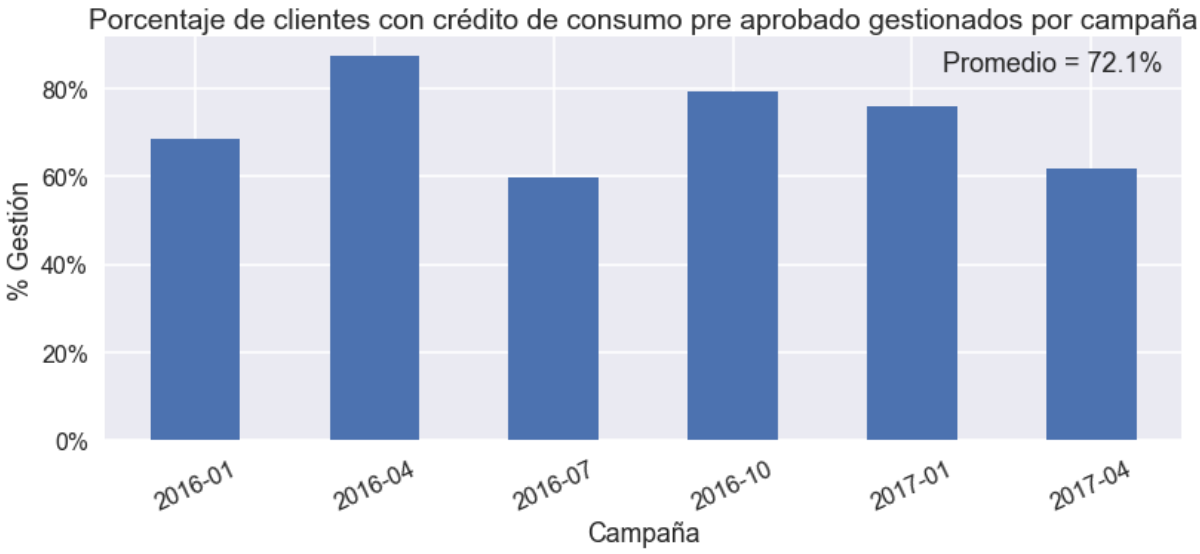
Ilustración 8: Resultados de ventas de créditos de consumo como porcentaje de clientes gestionados, no gestionados y total de clientes pre aprobados de las últimas 6 campañas.



Fuente: Elaboración propia.

En la Ilustración 9, se puede observar que en promedio se logra gestionar un 72% de los clientes, con un mínimo de 60% y un máximo de 87% de gestión, además, se tiene que la meta de los ejecutivos es gestionar el 80% de los clientes.

Ilustración 9: Porcentaje de clientes con crédito de consumo pre aprobado gestionados por campaña.



Fuente: Elaboración propia.

Por lo tanto, considerando que la gestión de clientes no es garantizada, ya que, en promedio un 28% de los clientes no son gestionados y la tasa de respuesta es considerablemente mayor cuando se gestiona al cliente, es importante realizar la gestión de clientes más propensos a aceptar el producto ofrecido y que no sea una pérdida de tiempo para los ejecutivos.

2.2 Oportunidades y consecuencias

Se pueden identificar tres oportunidades de mejora que afectan a la venta de colocaciones de consumo.

1. Se puede mejorar la asignación del producto que se va a ofrecer a cada cliente. Entender cuáles son las características que definen que un cliente necesite un crédito de consumo y/o un aumento de cupo en la tarjeta de crédito es fundamental para no hacer una oferta errada y aumentar la probabilidad de que el cliente tome la oferta.

2. Se puede mejorar la asignación de prioridades de los ejecutivos al momento de gestionar las campañas. El problema es que los ejecutivos son un recurso limitado y no logran gestionar el 100% de los clientes. En promedio, sólo el 70% de los clientes es gestionado y como se mencionó anteriormente, estos clientes poseen una tasa de respuesta mayor.

3. Se puede realizar una asignación exclusiva de los canales que se utilizarán para la promoción de los productos pre aprobados para cada cliente. Entender cuáles son los canales que el cliente utiliza para interactuar con el Banco es fundamental para hacer llegar la oferta al cliente por el canal adecuado, por ejemplo, si un cliente nunca lee los correos electrónicos debo asegurarme de utilizar otro canal para contactarlo.

La primera y segunda oportunidad pueden ser causadas porque se hace la asignación de prioridades y oferta en base a reglas básicas de comportamiento transaccional, de segmentación y tenencia de producto de los clientes, pero estas reglas no se encuentran respaldadas estadísticamente y no entregan los resultados esperados.

En la Tabla 2, se observan los resultados de ventas de créditos de consumo de la campaña del tercer trimestre del año 2017, se destacan las prioridades 1, 2, 3 y 4, ya que, estos corresponden a los clientes con mayor prioridad y mejor tasa de respuesta.

El detalle de cómo están construidas las prioridades es el siguiente:

- Prioridad 1: clientes que poseen una oferta que permite comprar la deuda que los clientes poseen en otras entidades financieras

- Prioridad 2: clientes a los que les falta pagar las últimas 6 o menos cuotas de un crédito de consumo

- Prioridad 3: clientes que hacen uso del 60% de la tarjeta de crédito

- Prioridad 4: clientes que poseen oferta de crédito de consumo y el ratio entre la deuda de consumo máxima de los últimos 3 meses y los últimos 12 meses es mayor al 80%

- Prioridad 5: clientes que poseen oferta de crédito de consumo y tarjeta de crédito

- Prioridad 6: clientes que poseen oferta de crédito de consumo o tarjeta de crédito

- Prioridad 7: clientes sin prioridad.

Si bien los grupos de prioridad 1 y 2 poseen un aumento considerable en la tasa de respuesta con respecto a las demás prioridades, estos conjuntos son muy pequeños y representan un 3,5% de los clientes. Además, en estas dos prioridades sólo se captura un 11,5% de la venta. El mismo análisis se puede realizar observando la **Tabla 27**, ubicada en anexos.

Tabla 2: Resultados de gestión y ventas de créditos de consumo por prioridad en campaña de julio, agosto y septiembre de 2017.

Prioridad	Porcentaje del total de ventas de c.c.	Tasa de respuesta de c.c.	Porcentaje de clientes con oferta de c.c.	Porcentaje de gestión
1	**1,5%**	**8,0%**	**0,6%**	**94,9%**
2	**10,5%**	**12,1%**	**2,9%**	**91,3%**
3	**9,3%**	**5,6%**	**5,4%**	**88,1%**
4	**26,0%**	**4,5%**	**19,1%**	**82,9%**
5	28,5%	2,4%	38,8%	80,0%
6	11,8%	4,2%	9,2%	76,7%
7	12,3%	1,7%	24,0%	54,7%
Total	**100%**	**3,3%**	**100%**	**75,1%**

Fuente: Elaboración propia.

- Por lo tanto, no se observa una buena priorización de los clientes, ya que, sólo el 47% de las ventas de créditos de consumo son capturadas en las primeras 4 prioridades y el 52,7% de las ventas se realiza en los tres grupos de menor prioridad, que corresponden al 72% de los clientes con oferta de crédito de consumo. Esto es respaldado en la tasa de respuestas de tarjetas de crédito, donde sólo un 35,6% de las ventas es realizada en las primeras 4 prioridades.

Por otro lado, considerando que en la campaña del segundo trimestre del año 2017 se realizó la gestión del 62% de los clientes (ver Tabla 3) y al utilizar el modelo de propensión a créditos de consumo desarrollado en este trabajo, se puede estimar que, si se hubiese gestionado al 60% de los clientes con mayor probabilidad de respuesta y al mismo tiempo la tasa de respuesta de los clientes gestionados no cambiase, entonces se aumentaría en 0,4 puntos porcentuales la tasa de respuesta general como

se muestra en la Tabla 3, lo que equivale a un 12% de aumento en las colocaciones de consumo pre aprobadas o el equivalente a MM$950. Cabe destacar, que en esta estimación no se toma en cuenta la tasa de respuesta de los clientes no gestionados, por lo tanto, es una estimación conservadora.

Tabla 3: Resultados de campaña del segundo trimestre del 2017. Clientes clasificados en deciles de probabilidad de respuesta, según modelo de propensión de créditos de consumo. Entre paréntesis se encuentra el caso hipotético de que se gestiona el 100% de los primeros 6 deciles y se mantiene la tasa de respuesta de estos deciles.

Decil	Porcentaje de gestión	Tasa de respuesta de clientes gestionados	Tasa de respuesta general
1	75% (100%)	16% (16%)	14% (16%)
2	71% (100%)	10% (10%)	8% (10%)
3	67% (100%)	5% (5%)	4% (5%)
4	66% (100%)	4% (4%)	3% (4%)
5	61% (100%)	2% (2%)	2% (2%)
6	61% (100%)	2% (2%)	2% (2%)
7	58% (0%)	1% (0%)	1% (0%)
8	55% (0%)	1% (0%)	1% (0%)
9	51% (0%)	1% (0%)	0% (0%)
10	52% (0%)	1% (0%)	1% (0%)
Total	**62% (60%)**	**4,8% (3,9%)**	**3,5% (3,9%)**

Fuente: Elaboración propia.

Se pone el foco en la priorización, debido a que, los ejecutivos logran gestionar a los clientes con mayor prioridad como se observa en la Tabla 2, pero no gestionan a los clientes con mayor probabilidad de compra, debido a la priorización que se les entrega.

- Por último, al integrar todas estas oportunidades se puede mejorar la tasa de respuesta de las campañas de promoción de productos de consumo pre aprobados y como consecuencia aumentar las colocaciones de consumo. Para esto, es necesario asignar las prioridades correctamente, de acuerdo con el producto que el cliente necesita, la oferta realizada, los canales que el cliente usa y la probabilidad de respuesta positiva del cliente, esto implica utilizar eficientemente a los ejecutivos para contactar a los clientes.

3 Objetivos

3.1 Objetivo general:

Desarrollar una metodología para aumentar la tasa de respuesta de las campañas de promoción de créditos de consumo y aumentos de cupo en tarjetas de crédito pre aprobadas y en consecuencia aumentar las colocaciones de consumo.

3.2 Objetivos específicos:

1. Construir dos modelos para predecir la probabilidad de respuesta positiva frente a las campañas de promoción de productos pre aprobados, tanto para créditos de consumo como para aumentos de cupo u obtención de tarjeta de crédito.

2. Segmentar a los clientes de acuerdo con el uso de canales: uso del sitio web y apertura de correos electrónicos.

3. Proponer la priorización de clientes en la gestión de las campañas, que considere la probabilidad de respuesta positiva a la campaña, el producto ofertado y el perfil de uso de canales del cliente (a partir de la segmentación del punto anterior).

4. Plantear un diseño experimental que permita validar el aumento de la tasa de respuestas de las campañas de promoción de productos pre aprobados producto de la priorización propuesta.

4 Alcances

Los productos de consumo pre aprobados que son considerados en el trabajo corresponden a los créditos de consumo pre aprobados y aumentos de cupo en tarjeta de crédito. Se excluye el aumento de cupo en línea de sobregiro, debido a que, representan menos del 10% de las ventas, además, sólo el 2% de los clientes posee oferta de línea de sobregiro en la campaña del tercer trimestre del año 2017, por lo que se considera que se cuenta con muy pocos datos y no es un producto relevante en comparación con los créditos de consumo y tarjetas de crédito.

En cuanto a la segmentación del uso de los canales por parte de los clientes, se hace en base al uso histórico de los canales, sin embargo, no se pretende cuantificar el impacto de cada canal en la tasa de respuesta (tasa de conversión del canal), debido a que, sería necesario realizar un experimento.

En este trabajo no se toma en consideración el canal SMS, debido a que, el uso de este canal se realiza desde agosto del año 2017, por lo tanto, se tienen muy pocos datos y este canal se usa en complementación a los clientes que no usan los canales de sitio web y correo electrónico, por lo tanto, la cantidad de datos es aún más reducida.

Sólo se plantea un diseño experimental que permita validar el aumento de la tasa de respuestas, no se lleva a cabo.

Estos experimentos no se realizan, debido a que, no se alcanzan a analizar los resultados de acuerdo con la planificación y los tiempos de las campañas.

5 Marco conceptual

A continuación, se presentan las herramientas que se utilizan en el desarrollo del trabajo.

5.1 Método de detección de valores atípicos

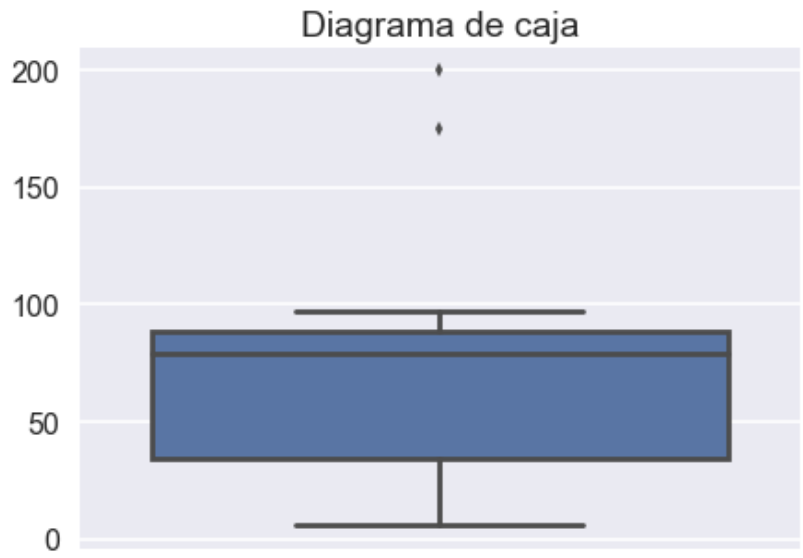
Para la detección de valores atípicos se utilizan los diagramas de cajas (boxplot), ya que su construcción define los valores atípicos según el rango inter-cuartil (Han, Pei, & Kamber, 2011).

Sea $x_1, x_2, \dots, x_n$ un conjunto de observaciones en orden creciente para una variable numérica. Los cuartiles (Q) dividen a la distribución de observaciones en cuatro intervalos de igual tamaño. El cuartil 1, 2 y 3 es al 25%, 50% y 75% respectivamente. La distancia entre el primer y tercer cuartil corresponde a una medida de dispersión que entrega el rango cubierto por el 50% de las observaciones. Esta distancia es llamada rango inter-cuartil (en inglés IQR).

En la Ilustración 10, se puede ver un ejemplo, en donde:

- Las líneas horizontales de arriba y abajo corresponden al mínimo y máximo del conjunto de observaciones que no son consideradas valores atípicos
- Los límites de la caja corresponden al primer y tercer cuartil
- La línea horizontal dentro de la caja corresponde a la mediana.

Ilustración 10: Ejemplo diagrama de caja



Fuente: Elaboración propia.

En caso de que las observaciones sean menores a Q1-1,5*IQR, se consideran valores atípicos y se grafican individualmente, análogamente los valores que superen a Q3+1,5*IQR, también se consideran valores atípicos.

5.2 Modelos de respuesta utilizados en marketing directo

Para cumplir con el primer objetivo, se debe predecir la probabilidad de que el cliente decida tomar la oferta de un producto determinado. La variable que se desea predecir es una variable binaria.

En la Tabla 4 se resumen los diferentes modelos cuantitativos utilizados en marketing directo para realizar esta tarea y según la división realizada por Indranil Bose y Xi Chen (Bose & Chen, 2009) estos se pueden clasificar en los que usan un enfoque estadístico (básico o avanzado) y en los que usan un enfoque de aprendizaje automático basado en minería de datos. Los modelos de respuesta han sido usados en la industria financiera y en el marketing directo según varios autores, por lo tanto, aplican para este problema (Ayetiran & Adeyemo, 2012) (Amini, Rezaeenour, & Hadavandi, 2015).

Tabla 4: Clasificación de modelos para respuesta binaria o probabilidad.

Clasificación	Técnica
Estadística básica	Logit/Probit
	Beta/Gamma
	Análisis discriminante
Estadística avanzada	Dos etapas Beta + Gamma
	Dos etapas Logit + Lineal
	Dos etapas Probit + Non-linear
	Modelo logit/probit de clase latente
Aprendizaje automático	Redes neuronales artificiales
	Árboles de decisión
	Naive Bayes
	SVM
	Random Forest

Fuente: Elaboración propia.

En el estudio realizado por Aaron Knoot, Andrew Hayes y Scott A. Neslin (Knott, Hayes, & Neslin, 2002) se demostró que al utilizar un modelo “next-product-to-buy” se mejoró considerablemente en comparación con un modelo heurístico y con un grupo de control. Ellos evaluaron 4 técnicas, análisis discriminante, regresión logística multinomial, regresión logística y redes neuronales. También estudiaron el impacto del tipo de información incluida en los modelos y el tipo de muestra para la calibración. Para esto, realizaron 88 modelos predictivos y ajustaron una regresión lineal sobre la exactitud de los modelos. Los resultados de la regresión lineal quedan resumidos en la Tabla 5, donde la variable Productos indica si el modelo incluye información de tenencia de

productos, la variable 0-1 toma el valor de 1 si la tenencia de productos fue codificada a una variable binaria y 0 si la tenencia de productos es considerada como la cantidad de productos, la variable Demo indica si el modelo incluye características demográficas, Valor indica si el modelo incluye el valor monetario del cliente, Muestra indica si el muestreo fue aleatorio, Disc indica si el modelo es análisis discriminante, Logit indica si el modelo es una regresión logística y Mnlogit indica si el modelo es una regresión logística multinomial.

Tabla 5: Regresión lineal de exactitud sobre tipo de datos del cliente, tipo de muestreo y técnica.

Categoría	Estadístico F	Variable	Coeficiente	Estadístico t	Significancia
Datos de cliente	77,7 (p<0,001)	Constante	41,66	80,1	0,000
		Productos	5,76	14,75	0,000
		0-1	0,51	1,45	0,152
		Demo	0,83	2,69	0,009
		Valor	1,83	5,93	0,000
Muestreo	392,0 (p<0.001)	Muestra	5,98	19,8	0,000
Técnica	2,13 (p=0.104)	Disc	-1,07	-2,5	0,014
		Logit	-0,66	-1,55	0,124
		Mnlogit	-0,56	-1,31	0,194

Fuente: Next-product-to-buy models for cross-selling applications (Knott, Hayes, & Neslin, 2002).

De los resultados obtenidos, red neuronal obtuvo resultados marginalmente mejores que regresión logística y regresión logística multinomial, sin embargo, el aumento no fue estadísticamente significativo mientras que la mejora por sobre análisis discriminante si fue estadísticamente significativa a un 95% de confianza.

Las técnicas basadas en la minería de datos han demostrado mejores resultados en el área bajo la curva ROC (métrica de evaluación, descrita más adelante) que las técnicas estadísticas (Coussement, Harrigan, & Benoit, 2015). En el estudio elaborado por Kristof Coussement, Paul Harrigan y Dries F. Benoit realizaron una clasificación en la cual el algoritmo CHAID es el mejor y le sigue CART, cabe destacar que los dos algoritmos son de árboles de decisión y que el peor algoritmo fue el de KNN-10 (K-Nearest Neighbor).

En conclusión, se utilizarán arboles de decisión y regresión logística, debido a que, son los modelos que en general han tenido los mejores resultados en la revisión bibliográfica y no existen diferencias significativas con respecto a las redes neuronales. Además, se puede validar los resultados del modelo al observar las reglas de decisión y los coeficientes de la regresión, por último, estos modelos presentan la gran ventaja de que son interpretables.

5.2.1 Árboles de decisión

Una de las grandes ventajas de los árboles de decisión (Linoff & Berry, 2004) es que representan reglas y se pueden interpretar fácilmente, lo que facilita su uso al tomar decisiones.

Un árbol de decisión es una estructura que se usa para dividir grandes cantidades de datos en pequeños grupos aplicando simples reglas de decisión. Mientras más subdivisiones posean, más y más parecidas son las entidades de cada subdivisión. Por lo tanto, el modelo de árboles de decisión consiste en un conjunto de reglas de decisión para dividir una población en grupos pequeños que tienen similitudes con respecto a una variable objetivo.

Los árboles de decisión pueden ser usados para clasificación binaria (1 o 0), estimar probabilidades o variables continuas.

Para la construcción de los árboles de decisión, los algoritmos buscan elegir la regla que mejor discrimine de acuerdo con la variable objetivo. Entonces, la primera tarea es elegir que variable independiente genera la mejor división. Para evaluar las divisiones se utiliza la pureza, baja pureza significa que el conjunto contiene una distribución representativa de las clases y alta pureza quiere decir que en el conjunto predominan los miembros de una clase. La mejor división es la que aumenta más la pureza de los nodos. Una buena división crea nodos de similares tamaños.

Para medir las diferencias de pureza, se utilizan diferentes test. Si la variable objetivo es categórica, se utilizan algoritmos que utilizan el test de impureza de Gini, ganancia de información o chi-cuadrado. Si la variable es continua, se utiliza el test de reducción de varianza o F-test.

La medida que utiliza la herramienta seleccionada para realizar la modelación corresponde a la impureza de Gini, donde para el nodo m con K posibles clases es definido de la siguiente forma:

$$Impureza\ de\ Gini = \sum_{k} p_{mk} * (1 - p_{mk})$$

Por otro lado, una de las reglas de parada que tiene relación con la impureza, es la reducción mínima de impureza, donde la reducción de impureza es definida de la siguiente forma:

$$Reducción\ de\ impureza = \frac{N_m}{N} * (impureza_m - \frac{N_{tR}}{N_m} * impureza_R - \frac{N_{tL}}{N_m} * impureza_L)$$

Donde,

- N es el número total de observaciones
- N_m es el número de observaciones en el nodo m
- N_{mL} (N_{mR}) es el número de observaciones en el nodo hijo izquierdo (derecho)

5.2.2 Regresión logística

Este modelo de estadística se utiliza para predecir el resultado de variables categóricas, donde se busca estimar la probabilidad de que una variable categórica tome un determinado valor dado ciertas variables explicativas. El modelo se puede escribir de la siguiente manera:

$$Logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta' * X_i$$

Luego, la probabilidad:

$$p_i = \frac{\exp(\beta * X_i)}{1 + \exp(\beta * X_i)}$$

Donde,

- X_i corresponde a las variables explicativas de la observación i
- β corresponde al vector de parámetros a estimar

La estimación de los parámetros se puede realizar a través del método de máxima verosimilitud. Los parámetros indican como afectan las variables explicativas en la probabilidad de ocurrencia. Este modelo es muy utilizado debido a su facilidad de interpretar.

5.3 Métodos de balanceo de clases

Es usual que en las campañas de marketing directo el número de respuestas positivas sea muy bajo, esto genera que los datos de entrenamiento sean muy desbalanceados y muchas veces esto es resuelto mediante técnicas de under sampling y over sampling (Amini, Rezaeenour, & Hadavandi, 2015).

Una de las técnicas más utilizadas en este tipo de problema es el random over sampling que consiste en aleatoriamente duplicar observaciones de la clase de menor frecuencia para obtener una muestra balanceada.

También se estudiará el desempeño de una técnica relativamente nueva y propuesta el año 2003, que consiste en realizar over sampling utilizando la técnica SMOTE y luego eliminar los enlaces de Tomek (Batista, Bazzan, & Monard, Balancing Training Data for Automated Annotation of Keywords: a Case Study., 2003) (Batista, Prati, & Monard, A study of the behavior of several methods for balancing machine learning training data, 2004).

La idea principal de SMOTE es crear nuevos ejemplos de la clase de menor proporción interpolando ejemplos cercanos.

Los enlaces de Tomek se definen de la siguiente manera:

- Dados 2 ejemplos, $\mathrm{E_i}$ y $\mathrm{E_j}$ pertenecientes a dos clases distintas y $\mathrm{d(E_i, E_j)}$ la distancia entre ellos.
- Un par $(\mathrm{E_i}, \mathrm{E_j})$ es llamado link de Tomek si no existe $\mathrm{E_l}$, tal que, $d(E_i, E_l) < d(E_i, E_j)$ o $d(E_j, E_l) < d(E_i, E_j)$.
- Si dos ejemplos son un link de Tomek, entonces alguno de ellos es ruido o los dos ejemplos están al borde de las clases.

5.4 Evaluación de los modelos de clasificación binaria

Existen diferentes métricas para la evaluación de los modelos de propensión (Bose & Chen, 2009), las siguientes métricas son las que se utilizarán y surgen a partir de la matriz de confusión, ver Tabla 6.

Tabla 6: Matriz de confusión

		Actual	
		Positiva	**Negativa**
Predicción	**Positiva**	A	B
	Negativa	C	D

Fuente: Elaboración propia.

La exactitud mide la proporción de observaciones correctamente clasificadas.

$$Exactitud\ o\ identificación\ correcta = \frac{A + D}{(A + B + C + D)}$$

La sensibilidad mide la proporción de observaciones positivas correctamente clasificadas como positivas.

$$Sensibilidad\ o\ ratio\ de\ respuesta\ = \frac{A}{(A+C)}$$

La especificidad mide la proporción de observaciones negativas correctamente clasificadas como negativas.

$$Especificidad\ = \frac{D}{(B+D)}$$

La precisión mide la proporción entre los verdaderos positivos y las observaciones clasificadas como positivas.

$$Precisión\ = \frac{A}{(A+B)}$$

La precisión mide la proporción entre los verdaderos negativos y las observaciones clasificadas como negativas.

$$Valor\ predictivo\ negativo\ o\ Negative\ Predictive\ Value\ (NPV)\ = \frac{D}{(C+D)}$$

El ratio de falsos positivos mide la proporción entre los falsos positivos y las observaciones negativas.

$$Ratio\ de\ falsos\ positivos\ = \frac{B}{(B+D)}$$

El F1 score es el promedio armónico de la precisión y la sensibilidad.

$$F1\ score = 2 * \frac{precisión * sensibilidad}{precisión + sensibilidad}$$

Finalmente, la última métrica de evaluación que se utiliza proviene de la curva ROC (Receiver Operating Characteristics).

Esta curva, se construye calculando la razón de sensibilidad y uno menos la razón de especificidad del modelo para distintos valores de calibración.

Como métrica de evaluación se utiliza el área bajo la curva ROC (en inglés Area Under Curve o AUC) que toma valores entre 0 y 1, siendo 1 el mejor resultado posible.

El modelo base con el cual se comparan los resultados es el modelo aleatorio que tiene un AUC de 0,5, por lo tanto, se espera que el modelo de predicción posea un AUC entre 0,5 y 1.

5.5 Segmentación de clientes

5.5.1 Variables RFM

Para cumplir con el segundo objetivo y realizar la segmentación de los clientes a través del algoritmo K-Means, el cual se detalla más adelante, se utilizan las variables RFM como input del método, con el objetivo de identificar segmentos de clientes con comportamientos de uso similares de la página web y correo electrónico.

Para agregar los datos transaccionales, también se utilizan las variables RFM como input de los modelos de propensión.

El modelo RFM ha sido aplicado en distintas áreas, pero especialmente para marketing directo e identificar a los clientes más valiosos. El modelo tradicional RFM realiza la segmentación dividiendo a la población de acuerdo con cada una de estas variables en 5 quintiles de igual tamaño (Wei, Lin, & Wu, 2010), sin embargo, en este caso se utiliza el algoritmo K-Means para determinar los grupos de clientes.

- “Recency” corresponde al periodo de tiempo desde la última compra
- “Frequency” corresponde al número de compras en un determinado periodo de tiempo
- “Monetary” corresponde a la cantidad de dinero gastado en un periodo de tiempo

Las variables “Frequency” y “Monetary” poseen una alta correlación, debido a que, mientras mayor sea la frecuencia de compras, mayor será el valor monetario que se gasta, por esto también se utiliza el monto promedio de gasto, como es propuesto por Charles Edmundson en el trabajo realizado por Claudio Marcus (Marcus, 1998).

En la siguiente tabla se pueden ver ejemplos de las variables que se utilizan para lograr capturar los distintos comportamientos de los clientes en el servicio web, correo electrónico, tarjeta de crédito y débito.

Tabla 7: Variables RFM para los distintos servicios.

Servicio	"Recency"	"Frequency"	"Monetary"
Sitio Web	Cantidad de días desde la última visita	Promedio de visitas mensuales	
E-Mail	Cantidad de días desde la última apertura	Porcentaje de correos abiertos	
Tarjeta de crédito y débito	Cantidad de días desde la última compra	Cantidad total de compras	Monto promedio de compras

Fuente: Elaboración propia.

5.5.2 Segmentación con algoritmo K-Means

Para cumplir con el objetivo de identificar a los clientes de acuerdo con el perfil de uso de los canales digitales, es necesario realizar una segmentación, este es un problema de clasificación no supervisado, debido a que, se deben encontrar los distintos segmentos sin tener una variable objetivo que se conozca su valor. Para esto se utilizará el algoritmo de K-Means.

El algoritmo de K-Means es utilizado para encontrar "K" segmentos definidos según la proximidad entre los datos. El algoritmo es el siguiente:

1. Se seleccionan k centroides aleatoriamente.
2. Se asigna cada una de las observaciones al centroide más cercano, generando k grupos de observaciones.
3. Para cada grupo, se calcula un nuevo centroide, definido por el promedio de las dimensiones de las observaciones de cada grupo.
4. Si el centroide se desplazó, entonces repetir el algoritmo desde el paso 2. Si no, el algoritmo ha finalizado.

Es importante destacar que la solución entregada por este algoritmo no siempre es el óptimo global y el resultado puede depender de los centroides utilizados al inicio. Por lo tanto, es recomendable partir el algoritmo varias veces con centroides distintos.

Finalmente, una gran desventaja de este algoritmo es que necesita como input la cantidad de grupos (k), sin embargo, a partir de la "regla del codo" se puede identificar el número correcto de grupos (Halkidi, Batistakis, & Vazirgiannis, 2001), que consiste en:

1. Correr el algoritmo de K-Means para todos los valores de k entre un $k_{mínimo}$ y un $k_{máximo}$.

2. Para cada valor de k, correr el algoritmo r veces con diferentes centroides de inicio.

3. Graficar los mejores valores de distancia total intra-cluster para cada valor de k como función de k.

4. Encontrar el k en que existe una diferencia local significativa en la distancia total intra-cluster. Esto se ve reflejado como un “codo” en el gráfico.

5.6 Métodos de selección de variables

5.6.1 Information Value

El Banco tiene a su disposición distintas bases de datos de las cuales es posible extraer una gran cantidad variables que son candidatas a ser buenos predictores para los modelos de propensión, en este trabajo se obtuvieron 476 variables, sin embargo, para entrenar los modelos se decide realizar una selección de variables, debido a que, entrenar los modelos con una gran cantidad de variables no es computacionalmente eficiente y se corre el riesgo de sobre ajuste.

Para hacer la selección de variables, se hace uso del valor de la información y el peso de la evidencia para estimar el poder predictivo de las distintas variables.

La fórmula para calcular el peso de la evidencia (Lin, 2013) (WOE, Weight Of Evidence) es la siguiente:

$$WOE_i = \left[\ln\left(\frac{\%\ clase\ positiva_i}{\%\ clase\ negativa_i}\right)\right]$$

Donde el subíndice i corresponde al intervalo o categoría de la variable a la que se desea calcular el peso de la evidencia.

Luego, el valor de la información (en inglés Information Value o IV) se calcula de la siguiente forma:

$$IV = \sum(\%\ clase\ positiva_i - \%\ clase\ negativa_i) x WOE_i$$

En la Tabla 8 se puede observar un ejemplo del cálculo del valor de la información y el peso de la evidencia, donde CP se refiere a Clase Positiva y CN a la Clase Negativa.

Tabla 8: Ejemplo de cálculo de valor de la información.

# compras	Total	# CP	# CN	% CP	% CN	WOE	IV
0	20.070	438	19.632	9,0%	20,6%	-0,8264	0,0959
1 a 3	21.260	596	20.664	12,3%	21,7%	-0,5697	0,0537
4 a 10	12.394	457	11.937	9,4%	12,5%	-0,2865	0,0090
11 a 25	8.744	350	8.393	7,2%	8,8%	-0,2010	0,0032
26 a 40	6.920	305	6.615	6,3%	7,0%	-0,1005	0,0007
41 a 60	5.613	314	5.299	6,5%	5,6%	0,1504	0,0014
61 a 100	9.886	651	9.235	13,4%	9,7%	0,3240	0,0120
100 a 200	7.234	575	6.659	11,9%	7,0%	0,5269	0,0256
más de 200	7.879	1.165	6.714	24,0%	7,1%	1,2248	0,2077
Total	**100.000**	**4.851**	**95.148**	**100%**	**100%**		**0,4092**

Fuente: Variable Reduction in SAS by Using Weight of Evidence and Information Value (Lin, 2013).

En este caso, es necesario realizar este cálculo para los dos modelos predictivos, en el que las clases positivas corresponden a la venta de créditos de consumo y tarjetas de crédito.

5.6.2 Recursive Feature Elimination

Una de las desventajas de aplicar un filtro a las variables, por ejemplo, Information Value, es que es un método que toma en cuenta sólo una variable a la vez, es decir, sólo toma en cuenta la utilidad de cada una de las variables con respecto a la variable objetivo.

A diferencia de los métodos de filtrado, los métodos wrapper, toman en cuenta el conjunto de variables y no cada una por sí sola, por lo tanto, esto permite seleccionar variables que en conjunto tienen mayor utilidad, en este caso se hace uso del método Recursive Feature Elimination (Guyon, Weston, Barnhill, & Vladimir, 2002), que consiste en:

1. Entrenar el clasificador
2. Calcular la importancia de cada variable
3. Eliminar las variables con menor importancia

5.7 Correlación entre variables

5.7.1 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida que denota la fuerza y dirección de la relación linear entre dos variables numéricas.

Esta medida toma valores entre -1 y 1, donde -1 denota una correlación linear negativa perfecta, 0 que no existe correlación y 1 que existe una correlación total positiva.

La fórmula para calcular el coeficiente de correlación de Pearson para una muestra es la siguiente:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Donde,

- n es el tamaño de la muestra
- $x_i(y_i)$ corresponde a la observación i de $x(y)$
- $\bar{x}(\bar{y})$ corresponde al promedio de las observaciones de $x(y)$

5.7.2 V de Cramér

Para medir la asociación entre variables categóricas es posible calcular el estadístico chi cuadrado, sin embargo, este estadístico depende del número de categorías que tiene cada variable, por esta razón, se utiliza el estadístico V de Cramér (Cramér, 1946) que es una medida estandarizada del grado de dependencia entre variables categóricas, esta toma valores entre 0 y 1, mientras mayor sea el valor del estadístico, mayor es la dependencia entre las variables.

La fórmula para calcular el estadístico es la siguiente:

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{(q-1)}}$$

Donde,

- n corresponde al número de observaciones

- q corresponde al mínimo entre el número de columnas y filas de la tabla de contingencia

5.8 Optimización

Dentro de la optimización están los problemas programación lineal. Estos se caracterizan porque la función objetivo y las restricciones son lineales.

Un problema de programación lineal generalizado tiene la siguiente forma (Bertsimas & Tsitsiklis, 1997):

$$
\begin{aligned}
&minimizar && c'x && \\
&sujeto\ a && a_i'x \geq b_i, && i\epsilon M_1, \\
& && a_i'x \leq b_i, && i\epsilon M_2, \\
& && a_i'x = b_i, && i\epsilon M_3, \\
& && x_j \geq 0, && i\epsilon N_1, \\
& && x_j \leq 0, && i\epsilon N_2,
\end{aligned}
$$

Donde,

- x_i corresponden a las variables de decisión
- c corresponde al vector de costos
- $A(B)$ corresponde a una matriz de son parámetros conocidos

Un vector x que satisface todas las restricciones se llama solución factible y un vector x que minimiza la función objetivo corresponde a una solución factible óptima.

Un caso particular es cuando algunas de las variables de decisión sólo toman valores enteros, este problema es llamado programación entera.

5.9 Diseño experimental

Para estudiar si una promoción tuvo el efecto deseado y validar la hipótesis de que se aumenta la tasa de respuesta, se pueden utilizar los experimentos.

En el diseño de experimentos, los factores corresponden a las variables que se desean testear y los niveles de los factores corresponden a los valores que pueden tomar estas variables.

Un diseño de experimentos simple corresponde al diseño factorial completo, que consiste en realizar todas las posibles combinaciones de factores. En este tipo de diseño el total de combinaciones aumenta rápidamente con el número de factores y niveles, por lo tanto, en algunos casos no es factible realizar todas las combinaciones. Por ejemplo, un experimento con dos factores y tres niveles en cada factor posee 64 combinaciones.

Otro tipo de diseños experimentales es el diseño factorial fraccional que disminuye el número de combinaciones que se testean al costo de que algunos efectos no se pueden de diferenciar de otros.

Para este caso, el factor que se desea estudiar es el tipo de priorización que se realiza al gestionar a los clientes, que puede ser la priorización actual o la planteada en esta memoria, por lo tanto, se tienen pocas combinaciones y es posible realizar un diseño factorial completo.

5.9.1 Grupo de control y de tratamiento

Con el objetivo de determinar que un factor influye en la variable dependiente y evitar conclusiones erróneas se necesita de un grupo de control y de tratamiento.

El grupo de control corresponde a los clientes que no reciben por ningún medio de comunicación la promoción y el grupo de tratamiento corresponde a los clientes que reciben la promoción.

5.9.2 Muestreo

Si se desea estimar una proporción de población, entonces para determinar el tamaño de una muestra estadísticamente significativa se tiene que:

$$n = \frac{\chi^2 * N * P * (1-P)}{d^2 * (N-1) + \chi^2 * P * (1-P)}$$ (Krejcie & Morgan, 1970)

Donde,

- n corresponde al tamaño de muestra necesario
- χ^2 corresponde al valor de la tabla chi-cuadrado para un grado de libertad al nivel de confianza deseado
- N corresponde al tamaño de la población
- P corresponde a la proporción de la población, sin embargo, se puede asumir que P es igual a 0,5, ya que, entrega el tamaño de muestra máximo
- d corresponde al error deseado

Se tiene que para N lo suficientemente grande, la fórmula es la siguiente:

$$n = \frac{\chi^2 * P * (1 - P)}{d^2}$$

5.9.3 Test de hipótesis

Una hipótesis es una afirmación acerca de una característica de una población y un test de hipótesis o test de significancia es un procedimiento para para comprobar una hipótesis (Triola, 2017).

Los test de hipótesis contienen una hipótesis nula (H_0) y una hipótesis alternativa (H_1). La hipótesis nula corresponde a la afirmación de que una característica de la población es igual a un determinado valor. La hipótesis alternativa es la afirmación de que la característica de la población es distinta al valor de la hipótesis nula.

Existen distintos tipos de test de hipótesis, entre ellos:

- Test de una (dos) proporción (es)
- Test de un (dos) promedio (s)
- Test de una (dos) desviación (es) estándar o varianza (as)

En este trabajo, se utiliza el test de dos proporciones, ya que, se desea estudiar la proporción de clientes que toma un crédito de consumo o un aumento de cupo.

5.9.3.1 Test de dos proporciones

Este test se utiliza para determinar la veracidad de una afirmación acerca de dos proporciones.

Los requisitos para que el test sea válido son los siguientes (Triola, 2017):

1. Las muestras son aleatorias.
2. Las muestras son independientes.
3. Cada una de las muestras debe tener al menos 5 casos positivos y 5 casos negativos.

Para $H_0: p_1 = p_2$ el test estadístico es el siguiente:

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\bar{p} * (1 - \bar{p}) * (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Donde,

- $p_i = proporción\ de\ la\ población\ i$
- $n_i = tamaño\ de\ la\ muestra\ i$
- $x_i = número\ de\ casos\ positivos\ en\ la\ muestra\ i$
- $\widehat{p_i} = \frac{x_i}{n_i}$
- $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$

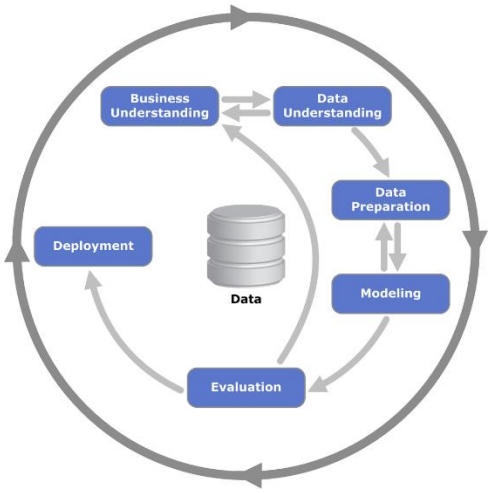
La hipótesis nula se rechaza si el valor $z > z_{\frac{\alpha}{2}}$, donde α corresponde al nivel de confianza deseado.

6 Metodología y desarrollo metodológico

Una variante de la metodología KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), CRISP-DM es la que se utiliza en este trabajo, el cual entrega una pauta para descubrir patrones de comportamiento o información útil en grandes bases de datos.

A grandes rasgos y como se observa en la Ilustración 11, la metodología consta de la comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (Shearer, 2000).

Ilustración 11: Metodología CRISP-DM.



Fuente: IBM SPSS Modeler CRISP-DM Guide

Esta metodología fue adaptada al caso y consta de la identificación de objetivos, recopilación de datos e información, preprocesamiento y limpieza de datos, transformación, selección de variables, minería de datos, interpretación y evaluación de los modelos de minería de datos, integración de los modelos desarrollados y diseño de experimento para validar el aumento en la tasa de respuesta.

1. La identificación de los objetivos radica en comprender la empresa, el área de la empresa en que se trabaja, la industria en que se inserta la empresa y las necesidades del negocio. Esta etapa se encuentra detallada en los puntos 1, 2, 3 y 4 de este informe.

2. La recopilación de datos consiste en elegir la información y conjuntos de datos que serán necesarios para cumplir con el objetivo del trabajo, separándola de la información innecesaria.

3. La etapa de preprocesamiento y limpieza de datos consiste en la identificación y tratamiento de datos faltantes y valores atípicos, en esta etapa también se considera un escalamiento de las variables para la elaboración del modelo de segmentación.

4. La fase de transformación de datos consiste en la creación de variables nuevas, a partir de las extraídas inicialmente, mientras que en la selección de variables se utiliza un filtro para eliminar variables que no tienen poder de predicción.

5. En la etapa de minería de datos se tiene en consideración la aplicación de los modelos de segmentación del uso de canales y los modelos de propensión a la compra de productos de consumo pre aprobados. Esta etapa considera la evaluación del poder predictivo de los modelos de propensión.

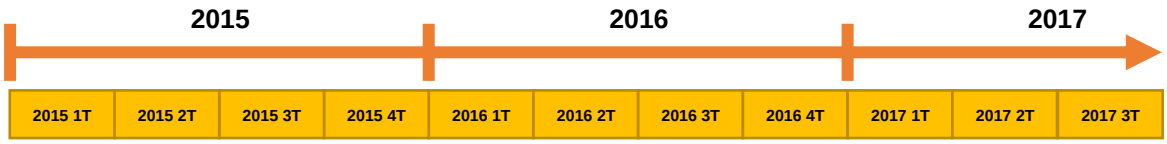
6. Luego, se resuelve un modelo de optimización que integra los modelos desarrollados anteriormente. El resultado de la optimización entrega el grupo de clientes a priorizar en la gestión de los ejecutivos, maximizando la utilidad esperada de este conjunto.

7. Finalmente, se plantea el diseño experimental para validar el aumento en la tasa de respuesta de los créditos de consumo y aumentos de cupo en tarjeta de crédito producto de la metodología diseñada.

6.1 Recopilación de datos e información

En esta etapa se analizan las bases de datos a las cuales se tiene acceso, como se puede observar en la Ilustración 12 estas corresponden a un horizonte de tiempo desde enero del 2015 a septiembre del 2017.

Ilustración 12: Esquema del horizonte temporal utilizado en la modelación.



Fuente: Elaboración propia.

Según el estudio realizado por Indranil Bose y Xi Chen (Bose & Chen, 2009) el tipo de información utilizado en los modelos cuantitativos de marketing directo se puede observar en la Tabla 9, en donde se detalla la importancia que representa la capacidad predictiva del tipo de información en el comportamiento de los clientes.

Tabla 9: Tipo de información usada en modelos cuantitativos, ejemplos e importancia.

Tipo de información	Ejemplo	Importancia
Información externa	Geográfica (dirección de residencia, trabajo y oficina), demográfica (edad, sexo, tamaño de familia, etc.), estilo de vida (hábitos, intereses, etc.) y socio gráfica.	Baja
Comportamiento interactivo del cliente con la empresa	Datos transaccionales (generalmente se utiliza RFM), feedback e historial de navegación (RF tiempo).	Alta
Características del producto.	Tamaño, color, precio, etc.	Alta
Características de la promoción/solicitud/canal	Diseño.	Baja

Fuente: Quantitative models for direct marketing: A review from systems perspective (Bose & Chen, 2009).

Por lo tanto, se selecciona información de los tipos anteriormente mencionados de las siguientes bases de datos:

CIF, Demográfico, Vehículos, Bienes Raíces, Activos, Pasivos, Tenencia de productos, Apertura de productos, Deuda en el Banco, Deuda en el sistema financiero informada por SBIF, Saldos vistas, Sitio web, Aplicación, Correos electrónicos, Call Center, Simulaciones de créditos pre aprobados, Transferencias electrónicas, Pagos automáticos cargados a la cuenta corriente y tarjeta de crédito, Tarjeta de débito, Tarjeta de crédito, Cupos y utilización de líneas de sobregiro, Cupo y deuda de tarjeta de crédito y Campañas de pre aprobados.

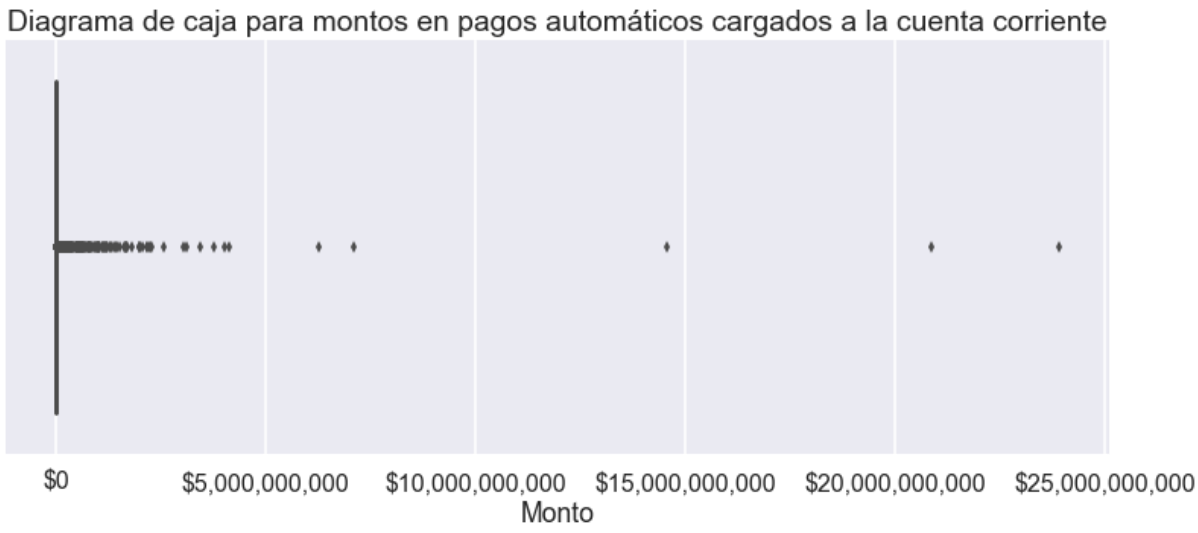
El listado de variables e información seleccionada se encuentra detallado en anexos, **Tabla 28**.

6.2 Preprocesamiento y limpieza de datos

Antes de realizar la transformación de variables, se realiza un análisis de valores atípicos, primero se grafica un diagrama de cajas para identificar la presencia de valores atípicos, sin embargo, al aplicar las reglas que fueron revisadas en la sección de marco conceptual de los diagramas de caja, se eliminan más del 5% de los datos, es por esta razón que se decide complementar el análisis utilizando valores máximos y mínimos.

Por ejemplo, para los montos de pagos automáticos cargados a la cuenta corriente se obtuvo el siguiente diagrama de cajas, este considera que aproximadamente el 10% de los datos corresponden a valores atípicos.

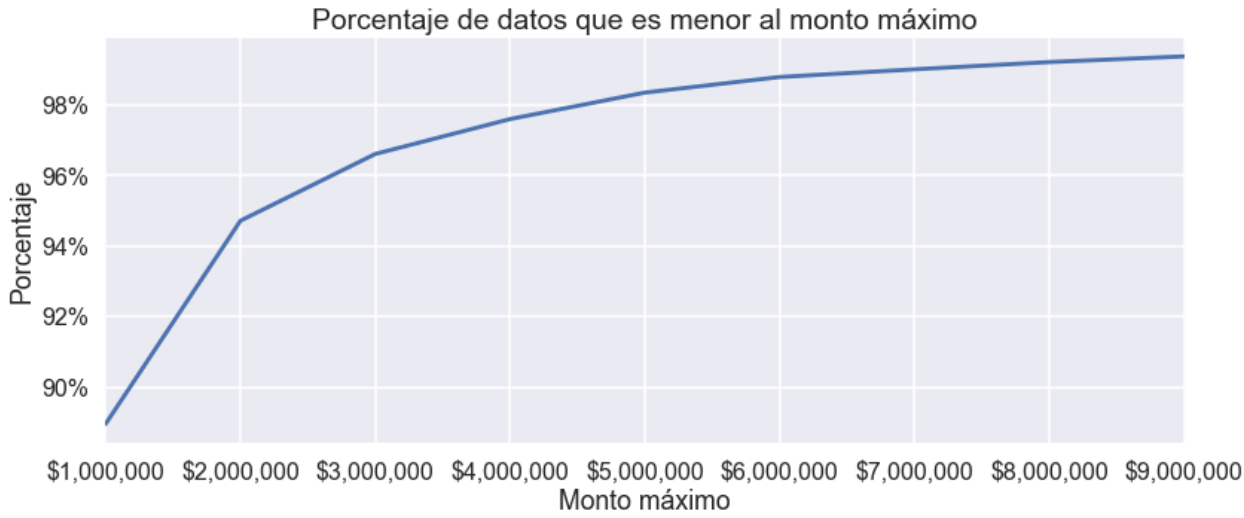
Ilustración 13: Resultado de diagrama de caja para montos en pagos automáticos cargados a la cuenta corriente.



Fuente: Elaboración propia.

Luego, al graficar el porcentaje de datos que es menor al monto máximo, se obtiene el siguiente gráfico.

Ilustración 14: Ejemplo para encontrar el monto máximo en pagos automáticos cargados a la cuenta corriente con el que no se consideran transacciones valores atípicos.



Fuente: Elaboración propia.

Por lo tanto, en este caso se decide eliminar los datos con compras mayores a $5.000.000, el procedimiento es análogo para encontrar el valor mínimo.

Como se puede observar en la Tabla 10, se eliminó aproximadamente el 2% de los datos. Los detalles de los resultados luego de filtrar los datos transaccionales se pueden observar en la **Tabla 29** ubicada en anexos.

Tabla 10: Resumen de eliminación de valores atípicos.

Base de datos	Porcentaje eliminado
Pagos automáticos cargados a cuenta corriente	1,8%
Pagos con débito	0,4%
Giros con débito	0,5%
Abonos TEF	3,3%
Cargos TEF	2,3%
Pagos automáticos cargados a tarjeta de crédito	1,7%
Pagos con tarjeta de crédito	2,3%

Fuente: Elaboración propia.

Con respecto a la imputación de missing values, se realiza la imputación de cero en los casos que corresponde, por ejemplo, en el número total de compras y monto total de compras con tarjeta de débito, debido a que, si el cliente no realiza compras en los periodos estudiados, entonces al realizar la creación de dichas variables se genera un missing value.

Por otro lado, se realiza la imputación de la mediana, ya que la distribución de este tipo de datos no es normal y no se tiene completitud, debido a que, los clientes sólo comprueban o actualizan los datos al solicitar un producto. El 4% de los montos de activos comprobados es imputado.

6.3 Transformación de datos

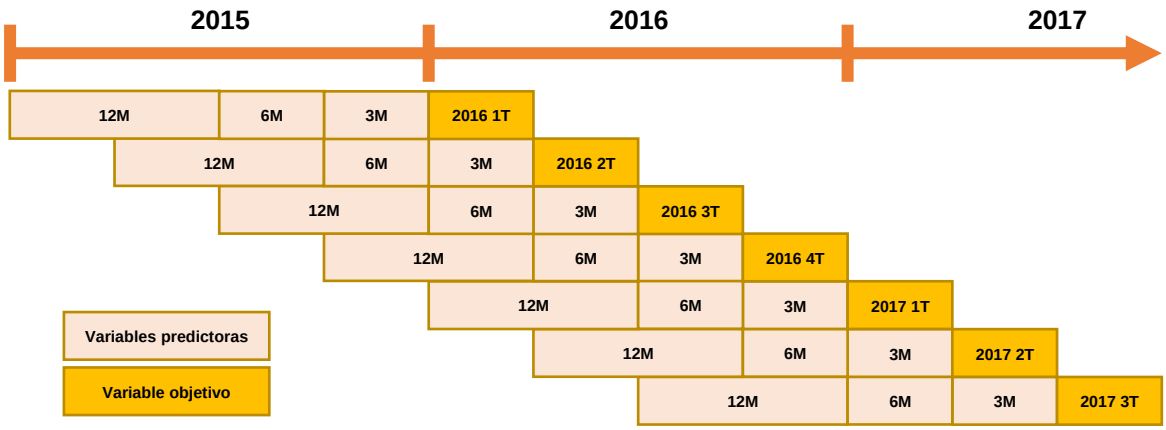
Al finalizar el tratamiento de valores atípicos, se definen tres periodos de tiempo en los cuales se analiza el comportamiento de los clientes con el fin de encontrar patrones de conducta en el corto, mediano y largo plazo que permitan predecir la respuesta de los clientes frente a las campañas.

Como se puede observar en la Ilustración 15, la variable a predecir u objetivo se encuentra definida desde el primer trimestre del año 2016 al tercer trimestre del año 2017, esta corresponde a un 1 si el cliente toma un crédito de consumo o 0 sino en el caso del modelo de propensión de crédito de consumo, análogamente se define la variable objetivo para el caso de tarjetas de crédito, donde la variable objetivo corresponde a 1 si el cliente toma un aumento de cupo o realiza la apertura de tarjeta de crédito con el monto ofertado o 0 si el cliente no toma la oferta.

Se toma la decisión de tomar en conjunto a los clientes que se les hace una oferta de apertura de tarjeta de crédito y a los clientes que se les hace una oferta de aumento de cupo en la tarjeta de crédito que tienen, ya que, sólo al 28% de los clientes se les hace una oferta de apertura de tarjeta de crédito y sólo el 0,2% de ellos toma la oferta, por lo

tanto, si se dividen estos clientes no se tendría una cantidad significativa de respuestas positivas en la apertura de tarjetas de crédito.

Ilustración 15: Esquema de la definición de periodos de tiempo para el cálculo de variables predictoras.



Fuente: Elaboración propia.

En el caso de las variables independientes, el primer periodo corresponde a los 3 últimos meses antes de las campañas, el segundo a los 6 últimos meses y el tercero a los últimos 12 meses.

Para las bases de datos en formato transaccional se utiliza el enfoque RFM mencionado en el marco conceptual y se calculan las siguientes variables:

1. Suma total
2. Máximo
3. Promedio
4. Promedio de días entre acción
5. Cantidad de días desde la última acción

Donde la acción puede ser una compra, transferencia, etc.

También se crearon variables que capturan el porcentaje de uso de la tarjeta de crédito en moneda nacional y extranjera. Para esto se divide el monto utilizado mensualmente por el cupo de la tarjeta de crédito.

Con respecto a las compras con tarjeta de crédito y débito, se capturó el rubro en el que más compras se hacen, el rubro en el que más dinero se gasta y cuanto se gasta en ese rubro.

Asimismo, se crearon variables binarias que capturan la tenencia de cuenta corriente en moneda nacional y extranjera, cuenta vista, tarjeta de débito, línea de sobregiro, depósitos a plazo, fondos mutuos y la cantidad de productos que el cliente posee.

Finalmente, para las variables numéricas de suma total, promedios y máximos se calculan dos ratios, el primero entre el periodo uno y dos, el segundo entre el uno y tres, con el objetivo de capturar cambios en el comportamiento de los clientes, por ejemplo, si un cliente aumenta el uso de la tarjeta de crédito esto se verá reflejado con una razón mayor a 1, debido a que, el promedio de las compras de los últimos tres meses debiese ser mayor al promedio de los últimos 12 meses.

Luego de realizar la transformación de los datos, se obtuvo un total de 476 variables en total. Por lo tanto, se hace necesario realizar una selección de las variables más importantes para cada modelo de propensión.

6.4 Selección de variables

En primer lugar, se realiza la discretización de las variables continuas de acuerdo con los quintiles, para luego calcular el valor de la información que se detalló en el marco conceptual.

Luego de realizar este procedimiento con todas las variables y tomando como clase positiva a la venta de créditos de consumo y aumentos de cupo o apertura en tarjeta de crédito, se ordenan las variables de acuerdo con el valor de información de cada una. Un extracto de las variables ordenadas de acuerdo con su poder predictivo con respecto a la venta de créditos de consumo se puede observar en la Tabla 11.

Tabla 11: Extracto del valor de información de las variables con mayor poder predictivo con respecto a la venta de créditos de consumo.

Variable	IV
Deuda de consumo promedio de los últimos 12 meses en el Banco	0,82
Deuda de consumo máxima de los últimos 12 meses en el Banco	0,79
Deuda de consumo promedio de los últimos 6 meses en el Banco	0,76
Deuda de consumo máxima de los últimos 6 meses en el Banco	0,75
Deuda de consumo promedio de los últimos 3 meses en el Banco	0,72
Deuda de consumo máxima de los últimos 3 meses en el Banco	0,71
Tenencia de crédito de consumo en el último mes	0,55
Días desde la última apertura de crédito de consumo en los últimos 12 meses	0,54
Número total de aperturas de créditos de consumo en los últimos 12 meses	0,54
Cantidad de días en los que realiza simulaciones de créditos de consumo en los últimos 12 meses	0,50
Cantidad de días desde que realizó última simulación de créditos de consumo en los últimos 12 meses	0,49
Cantidad total de simulaciones de créditos de consumo en los últimos 12 meses	0,49
Deuda de consumo promedio de los últimos 12 meses en el sistema financiero	0,47

Fuente: Elaboración propia.

Tomando en consideración que mientras mayor sea el valor de la información, mayor poder predictivo posee la variable, se eliminaron las variables que tenían un valor de la información menor a 10% procurando no eliminar ninguna de las variables que el área de inteligencia de negocios considera importante según su juicio experto.

Como se puede observar en la Tabla 11, a simple vista, muchas de las variables están correlacionadas, por lo que se calcula el coeficiente de correlación de Pearson entre las variables numéricas y el coeficiente de contingencia V de Cramér entre las variables categóricas con el objetivo de eliminar variables correlacionadas.

Luego se procede a seleccionar en orden las variables con mayor poder predictivo y eliminar las variables correlacionadas con esta. Por ejemplo, en la primera iteración se selecciona la variable “Deuda de consumo promedio de los últimos 12 meses en el Banco” y se eliminan las variables “Deuda de consumo máxima de los últimos 12 meses en el Banco”, “Deuda de consumo promedio de los últimos 6 meses en el Banco”, “Deuda de consumo máxima de los últimos 6 meses en el Banco”, “Deuda de consumo promedio de los últimos 3 meses en el Banco” y “Deuda de consumo máxima de los últimos 3 meses en el Banco” que tienen una alta correlación de Pearson.

Con el objetivo de no seleccionar más de 40 variables se define que una variable numérica con una correlación de Pearson mayor a 0,25 se elimina y en el caso de las

variables categóricas se eliminan las variables que tengan una V de Cramér mayor a 0,2.

Se seleccionan un total de 29 variables para el modelo de propensión de compra de créditos de consumo y 29 variables para el modelo de aumentos de tarjetas de créditos.

Para reducir aún más la cantidad de variables se utiliza el método de Recursive Feature Elimination, eliminando una variable en cada iteración y observando la métrica de AUC.

Finalmente, para el modelo de propensión de compra de créditos de consumo, un conjunto de 11 variables logró aproximadamente los mismos resultados en cuanto a AUC que el modelo entrenado con las 30 variables y para el modelo de propensión de aumentos de cupo en tarjetas de crédito se seleccionó un conjunto de 8 variables. En la **Tabla 30** y **Tabla 31** ubicadas en anexos, se pueden observar las variables seleccionadas para cada modelo.

6.5 Análisis exploratorio

Para entender el comportamiento de los clientes pre aprobados se realiza un análisis exploratorio de las variables seleccionadas para cada uno de los modelos de propensión.

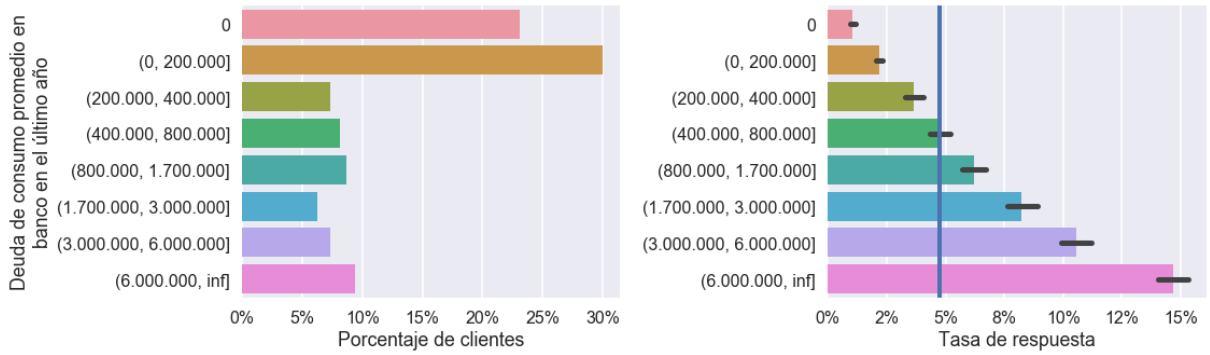
En las siguientes ilustraciones, en la figura de la izquierda se observa el porcentaje de clientes que cumple con la condición del eje y, mientras que en la figura de la derecha se observa la tasa de respuesta de cada condición del eje y. También se observa la tasa de respuesta promedio representada por una línea vertical azul en la figura de la derecha.

La tasa de respuesta promedio de la toma de créditos de consumo de clientes pre aprobados corresponde a 4,7% y la tasa de respuesta promedio de la toma de aumentos de cupo o apertura de tarjeta de crédito corresponde a 7,3%.

6.5.1 Variables utilizadas en modelo de créditos de consumo

En la Ilustración 16, se observa una de las variables más relevantes con respecto a la venta de créditos de consumo. En esta ilustración se observa cómo el 53% de los clientes pre aprobados poseen una deuda igual a 0 o menor a $200.000 con el Banco y que la tasa de respuesta aumenta con el aumento de la deuda de consumo en el Banco, es importante destacar que a partir de una deuda mayor a $800.000 la tasa de respuesta es mayor a la tasa de respuesta promedio.

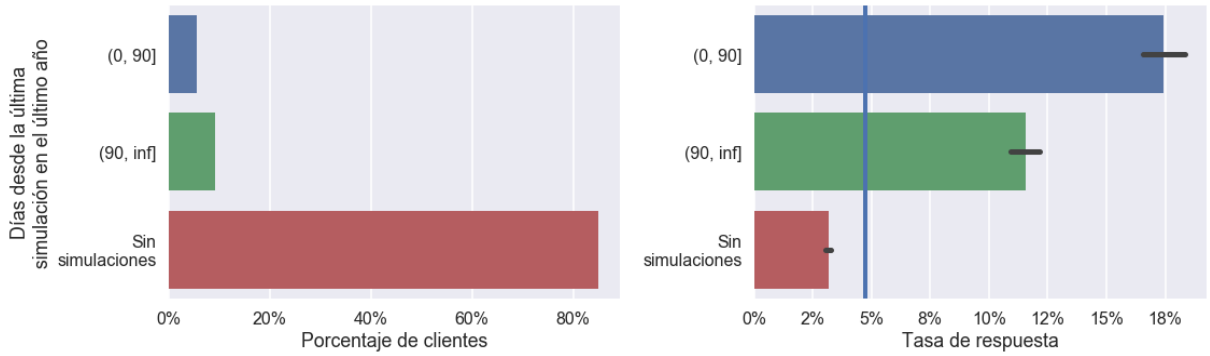
Ilustración 16: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo de deuda de consumo en Banco en el último año.



Fuente: Elaboración propia.

Otra de las variables más importantes corresponde a los días desde la última simulación de crédito de consumo, de la Ilustración 17 se puede observar que los clientes que realizan una simulación antes de la campaña poseen una alta tasa de respuesta y superior al promedio, a diferencia de los clientes que no realizan una simulación de crédito de consumo en el último año, ya que, estos clientes poseen una tasa de respuesta menor al promedio. Sin embargo, el porcentaje de clientes que realiza una simulación de crédito de consumo corresponde sólo al 15%.

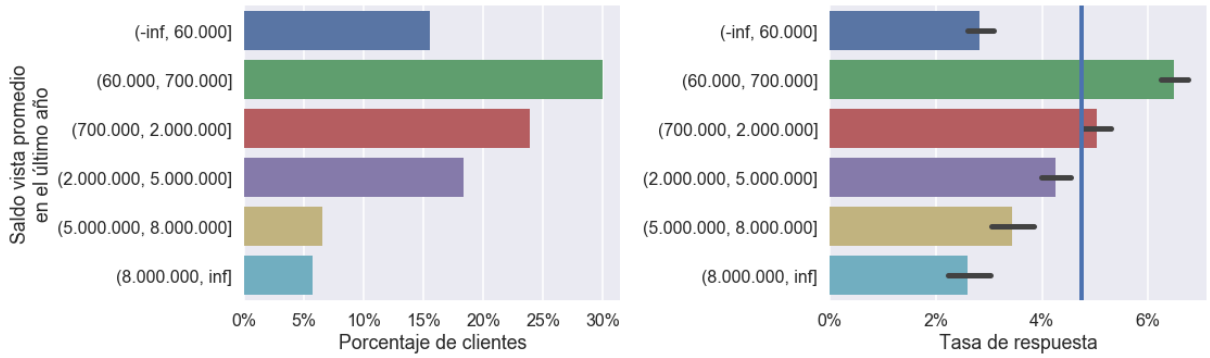
Ilustración 17: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo de días desde la última simulación de crédito de consumo en el último año.



Fuente: Elaboración propia.

La siguiente variable, es relevante para identificar a los clientes que no toman crédito de consumo. En la Ilustración 17 se observa que clientes con saldo vista promedio inferior a $60.000 poseen una tasa de respuesta bajo el promedio, sin embargo, en el rango entre $60.000 y $700.000 se posee una tasa de respuesta mayor al promedio. La tasa de respuesta decae a medida que el saldo vista promedio aumenta por sobre los $700.000. Por otro lado, el 42% de los clientes posee un saldo vista promedio entre $700.000 y $5.000.000, lo que da a entender que una parte importante de los clientes pre aprobados posee un alto poder adquisitivo.

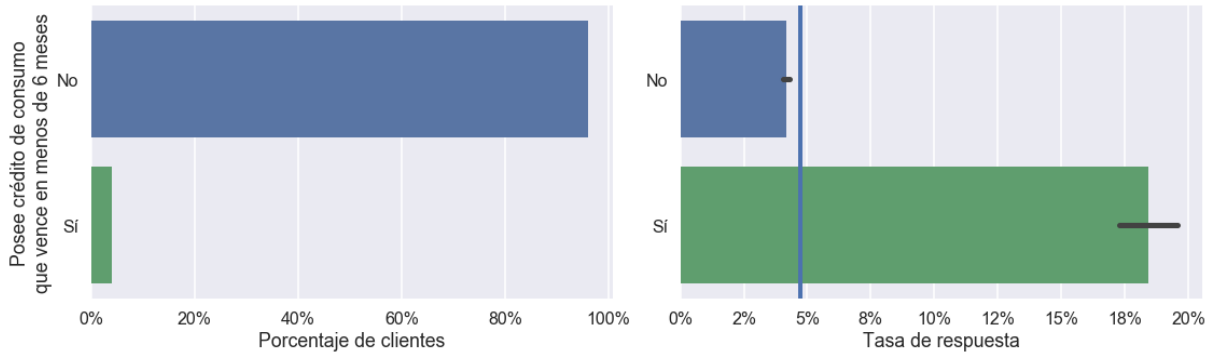
Ilustración 18: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo de saldo vista promedio en el último año.



Fuente: Elaboración propia.

La variable más usada por el área de inteligencia de negocios indica si el cliente está pagando las últimas 6 o menos cuotas de algún crédito. Como se puede observar en la Ilustración 19, la tasa de respuesta de estos clientes es de un 18,4%, muy superior al promedio, sin embargo, sólo el 4% de los clientes pre aprobados se encuentra en esta situación.

Ilustración 19: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tenencia de crédito de consumo que vence en menos de 6 meses.



Fuente: Elaboración propia.

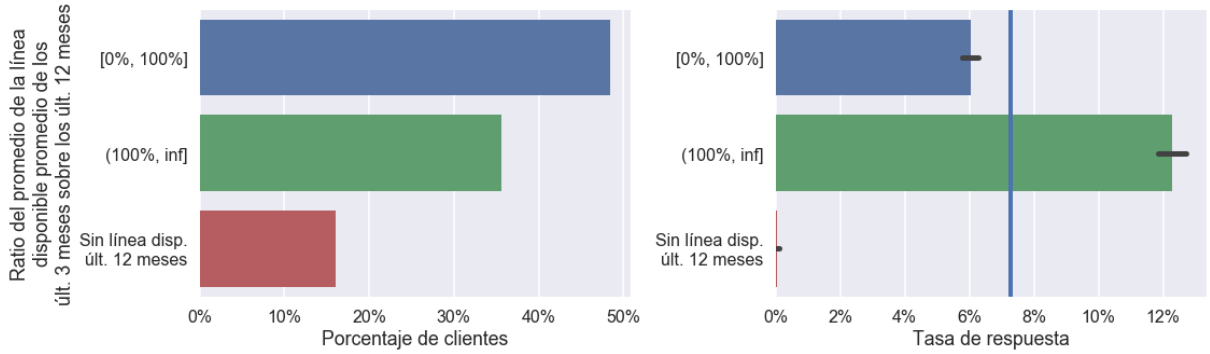
El resto de las variables se encuentra en anexos en la **Ilustración 42**, **Ilustración 43**, **Ilustración 44**, **Ilustración 45**, **Ilustración 46**, **Ilustración 47** e **Ilustración 48**. La interpretación es similar.

6.5.2 Variables utilizadas en modelo de propensión de tarjetas de crédito

Como se puede observar en la Ilustración 20, el 16% de los clientes pre aprobados no poseen línea de crédito disponible en los últimos 12 meses y tienen una tasa de respuesta casi nula.

Por otro lado, los clientes que han disminuido su línea de crédito poseen una tasa de respuesta menor a la del promedio, a diferencia de los clientes que han aumentado su línea de crédito con el banco, estos clientes poseen una tasa de respuesta mayor a la del promedio.

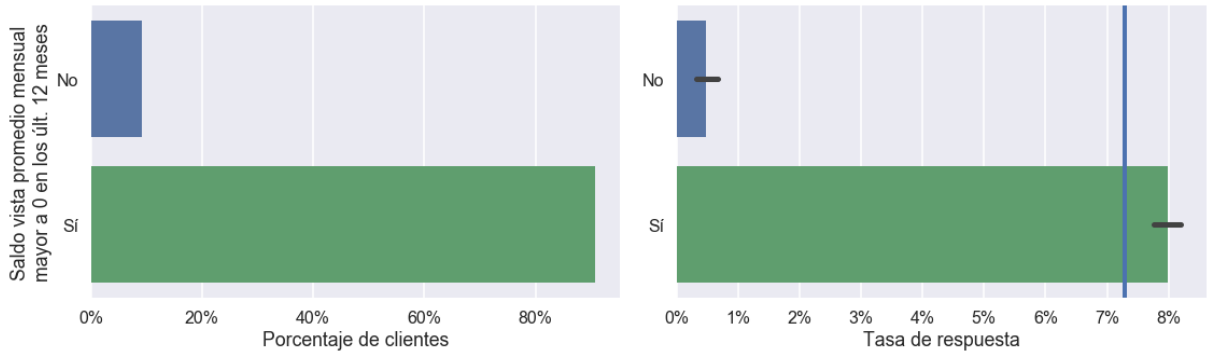
Ilustración 20: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por ratio del promedio de la línea de crédito disponible en el banco de los últimos 3 meses sobre los últimos 12 meses.



Fuente: Elaboración propia.

El saldo vista promedio mensual de los últimos 12 meses, no sirve para identificar a los clientes que toman la promoción de aumento de cupo, sin embargo, sirve para identificar a los que no la toman. Como se puede observar en la Ilustración 21, el 9% de los clientes no tienen saldo vista en los últimos 12 meses y su tasa de respuesta es casi nula.

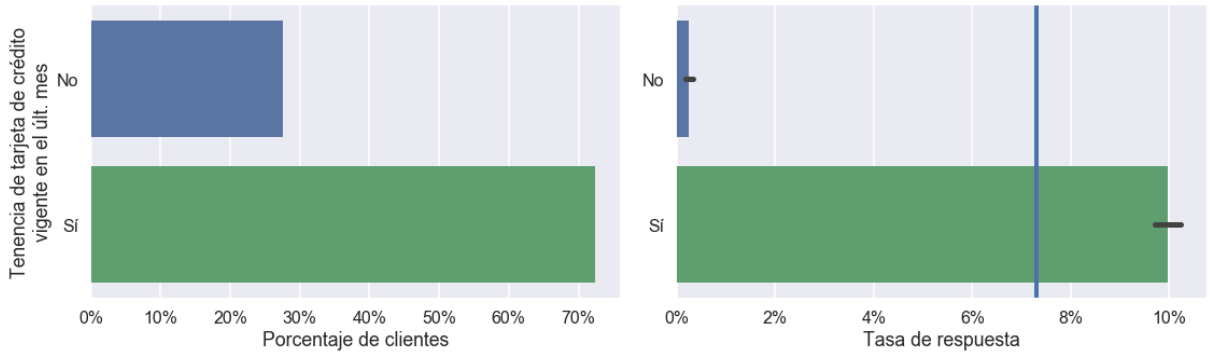
Ilustración 21: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por tenencia de saldo vista mensual promedio mayor a 0 en los últimos 12 meses.



Fuente: Elaboración propia.

La siguiente variable, también es un buen filtro para identificar a los clientes que no toman el aumento de tarjeta de crédito. En la Ilustración 22, se observa que el 28% de los clientes que no tienen una tarjeta de crédito vigente no toma el aumento de tarjeta de crédito pre aprobado.

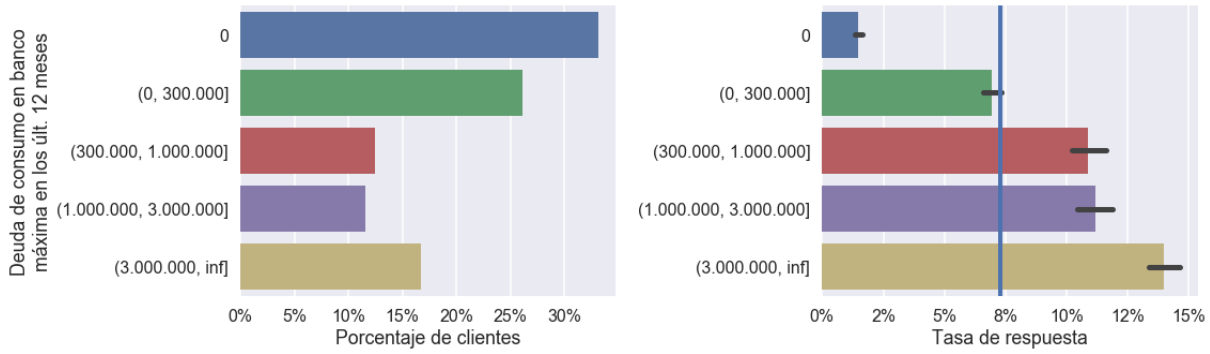
Ilustración 22: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por tenencia de tarjeta de crédito vigente.



Fuente: Elaboración propia.

La siguiente variable da a conocer quiénes son los clientes más propensos a tomar la oferta de aumento de tarjeta de crédito. En la Ilustración 23, se observa que los clientes que posee una deuda de consumo en el banco máxima en los últimos 12 meses mayor a $300.000 poseen una tasa de respuesta mayor al promedio.

Ilustración 23: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por tramo del máximo de la deuda de consumo en Banco en los últimos 12 meses.



Fuente: Elaboración propia.

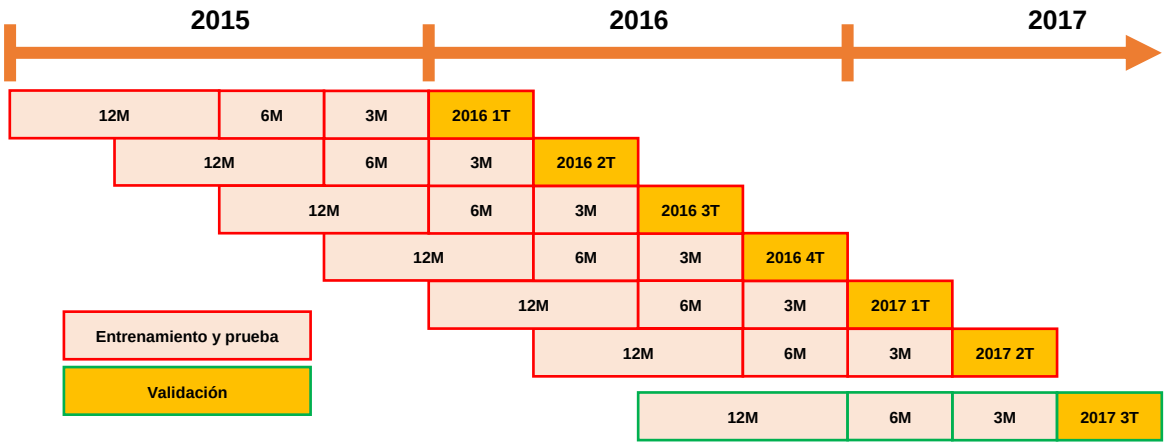
El resto de las variables se encuentra en anexos en la **Ilustración 49**, **Ilustración 50** e **Ilustración 51**. La interpretación es similar.

6.6 Minería de datos

6.6.1 Modelos de propensión

Con el objetivo de no entrenar modelos sobre ajustados a los datos de entrenamiento, se utilizan las campañas desde el primer trimestre del año 2016 al segundo trimestre del año 2017 con una muestra aleatoria de un 80% para el entrenamiento y un 20% para la prueba. Para la validación de los modelos se utilizan los datos de la campaña del tercer trimestre. Esto se puede ver gráficamente en la Ilustración 24.

Ilustración 24: Esquema de los datos utilizados para el entrenamiento, prueba y validación.

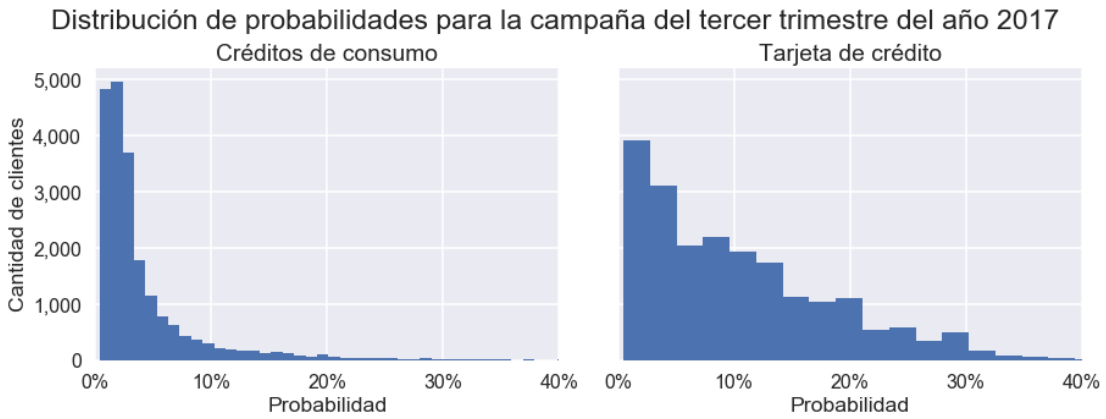


Fuente: Elaboración propia.

Como se mencionó anteriormente en el marco conceptual, este tipo de problemas usualmente presenta clases muy desbalanceadas, en el caso del modelo de créditos de consumo, la clase positiva representa un 4,7% de los datos tomando en cuenta todas las campañas, es por esta razón que, sólo para el caso de los árboles de decisión se estudian los métodos de Random Over Sampling y SMOTE combinado con la eliminación de enlaces de Tomek, ya que, en el caso de la regresión logística, el intercepto se hace cargo de este problema.

Para realizar la predicción de las clases, los modelos por defecto utilizan que las observaciones con una probabilidad mayor o igual al 50% pertenecen a la clase positiva, sin embargo, en el caso en que no se utilizan métodos de balanceo de clases, las probabilidades son mucho menores (ver Ilustración 25), por lo tanto, los modelos entregan que el 100% de las observaciones pertenecen a la clase negativa. Para solucionar este problema, se realiza una búsqueda de la probabilidad de corte que entregue el mejor resultado en cuanto a F1-Score, este indicador es el promedio armónico entre la precisión y la sensibilidad el cual es detallado en el marco conceptual.

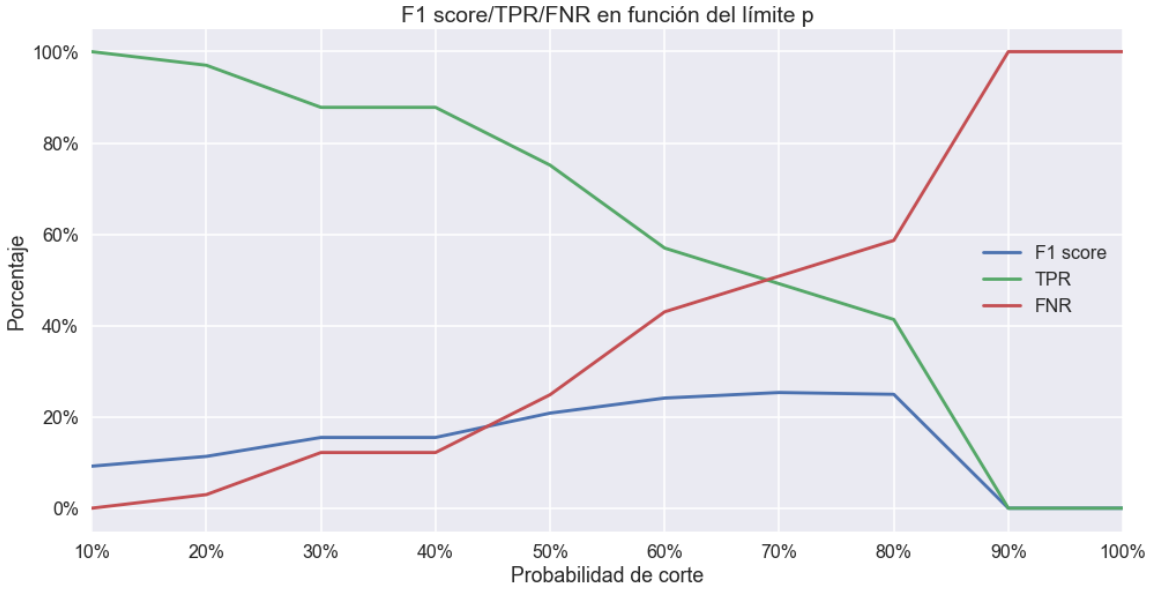
Ilustración 25: Distribución de probabilidades para la campaña del tercer trimestre del año 2017.



Fuente: Elaboración propia.

Por ejemplo, en la Ilustración 26, se observa que el mejor F1 score se obtiene con un corte de probabilidad de 70%, este balancea el porcentaje de verdaderos positivos y el de falsos negativos.

Ilustración 26: F1 score, Porcentaje de verdaderos positivos (TPR) y Porcentaje de falsos negativos (FNR) en función de la probabilidad de corte para el modelo de propensión de compra de créditos de consumo, árbol de decisión con método de balanceo Random Over Sampling.



Fuente: Elaboración propia.

En este problema en particular, se eligen dos métricas de evaluación como las más importantes.

En primer lugar, la sensibilidad de la clase positiva, es decir, que porcentaje de las ventas es capaz de capturar el modelo. La razón es que el beneficio de capturar una

venta es mucho más alto que el costo de realizar la gestión de un cliente que no toma la oferta.

En segundo lugar, el área bajo la curva ROC, debido a que, este indicador es una medida que permite evaluar que tan bien el modelo puede ordenar a los clientes, lo cual es un aspecto fundamental para realizar una buena optimización.

Lamentablemente, la exactitud, precisión, sensibilidad y F1-Score no son directamente comparables entre los modelos y no se puede decir que un modelo es mejor sólo en base a estos indicadores, ya que, estos dependen de la probabilidad de corte que se defina para realizar la predicción y como se mencionó anteriormente se define maximizando el F1-Score.

Por otro lado, no es posible definir una probabilidad de corte que permita realizar la comparación de los árboles de decisión, debido a que, la distribución de probabilidades depende de la cantidad de hojas y reglas utilizadas en cada árbol, por lo tanto, la única comparación que se puede hacer directamente es la de AUC, ya que, esta no depende de la probabilidad de corte.

6.6.1.1 Modelo de propensión de compra de créditos de consumo pre aprobados

Arboles de decisión

Los mejores resultados de los árboles de decisión con respecto al AUC obtenidos en la muestra de prueba, se entrenaron con los parámetros observados en la Tabla 12.

Tabla 12: Parámetros de parada para cada uno de los árboles de decisión. Modelo de propensión de créditos de consumo.

	Árbol de decisión		
Balanceo	**-**	**ROS**	**SMOTE + Tomek**
Máxima profundidad	6	6	6
Mínimo de observaciones en nodo padre	5%	-	-
Mínimo de observaciones en nodo hijo	3%	-	-
Reducción mínima de impureza[1]	-	0,2%	0,2%

Fuente: Elaboración propia.

La reducción de impureza mínima no se utiliza en el árbol de decisión sin balanceo, debido a que, este parámetro resulta ser muy sensible con clases desbalanceadas lo que dificulta la búsqueda del valor óptimo.

[1] Ver fórmula en el marco conceptual.

Los resultados de las métricas de desempeño para cada uno de los modelos de propensión a la toma de créditos de consumo se encuentran en la Tabla 13, donde la probabilidad de corte se define maximizando el F1-Score.

Tabla 13: Métricas de evaluación de árbol de decisión para los distintos métodos de balanceo de clases y regresión logística. Modelo de propensión de créditos de consumo.

	Árbol de decisión			Regresión logística
Balanceo	**-**	**ROS**	**SMOTE + Tomek**	**-**
Exactitud	**0,88**	0,86	0,87	0,87
NPV	**0,97**	**0,97**	**0,97**	**0,97**
Precisión	**0,18**	0,17	**0,18**	**0,18**
Especificidad	**0,90**	0,88	0,89	0,89
Sensibilidad	0,41	0,49	0,47	**0,50**
F1-Score	0,25	0,25	0,26	**0,27**
AUC	0,79	0,79	0,78	**0,81**

Fuente: Elaboración propia.

Para los árboles de decisión, en este caso en particular, sin realizar balanceo de clases y mediante el Random Over Sampling se obtienen los mejores resultados con respecto a el área bajo la curva de ganancia, logrando un aumento de 1 punto porcentual sobre el árbol con SMOTE + TOMEK.

Es importante destacar, que, mediante el uso de las técnicas de balanceo de datos, se logra aumentar la sensibilidad de la clase positiva, la métrica más importante. Mediante SMOTE + Tomek se logra aumentar en 6 puntos porcentuales la sensibilidad de la clase positiva y mediante Random Over Sampling se obtienen 2 puntos porcentuales adicionales en comparación con ROS.

Como se puede observar en la Ilustración 27, las variables que fueron seleccionadas por el mejor árbol de decisión (con ROS) corresponden a si el cliente ha realizado aperturas de créditos de consumo en el último año, si ha realizado simulaciones de créditos de consumo, el saldo vista mensual promedio de los últimos 12 meses y el monto de la deuda de consumo promedio en Banco de los últimos 12 meses.

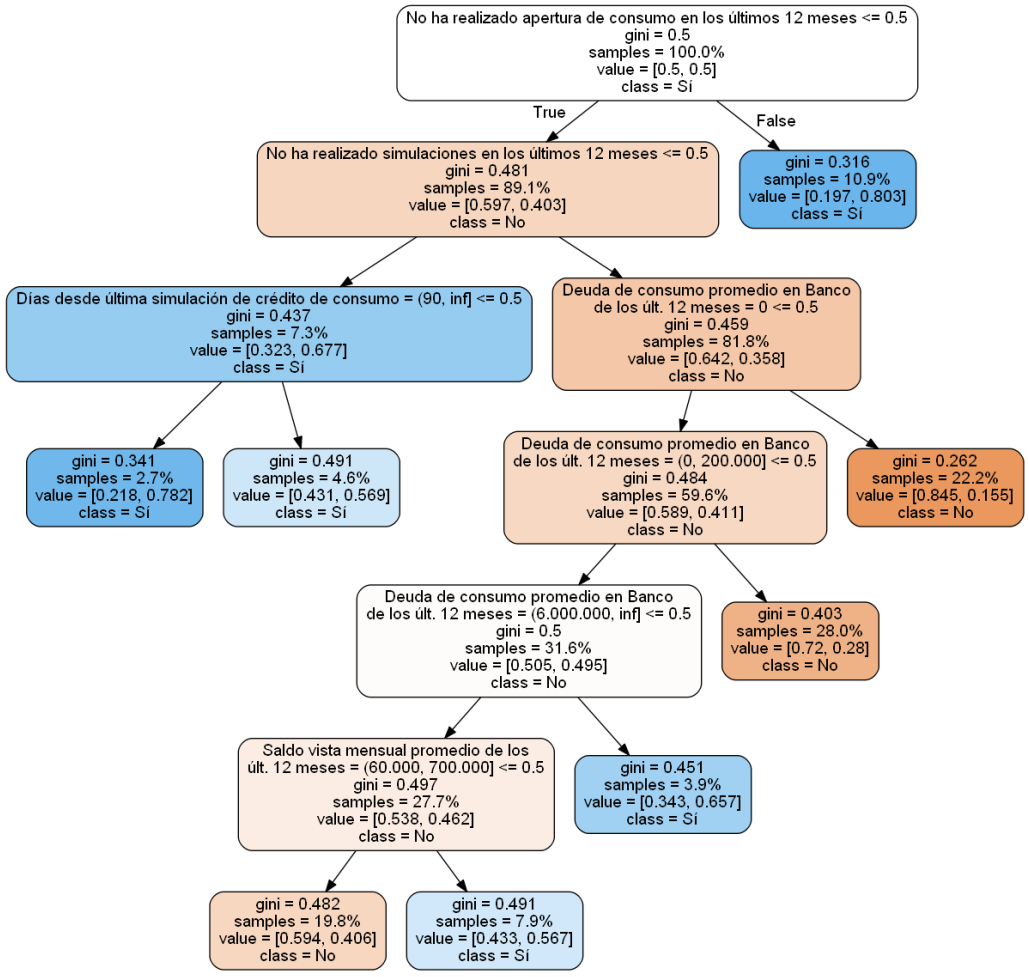
El árbol de decisión se puede leer de la siguiente forma:

- Los nodos hijos de la izquierda corresponden a que la condición del nodo padre es verdadera y análogamente para los nodos hijos de la derecha.
- Dentro de cada nodo se encuentra:
    - gini: El valor del coeficiente impureza de Gini.

- samples: El porcentaje de observaciones en el nodo
- value: El porcentaje de observaciones negativas y positivas respectivamente
- class: La clase dominante en el nodo

Para los arboles de decisión con over sampling, estos valores se encuentran alterados, debido a que, las proporciones y la cantidad de observaciones cambia al balancear las clases.

Ilustración 27: Árbol de decisión de modelo de propensión de compra de créditos de consumo con método Random Over Sampling.



Fuente: Elaboración propia.

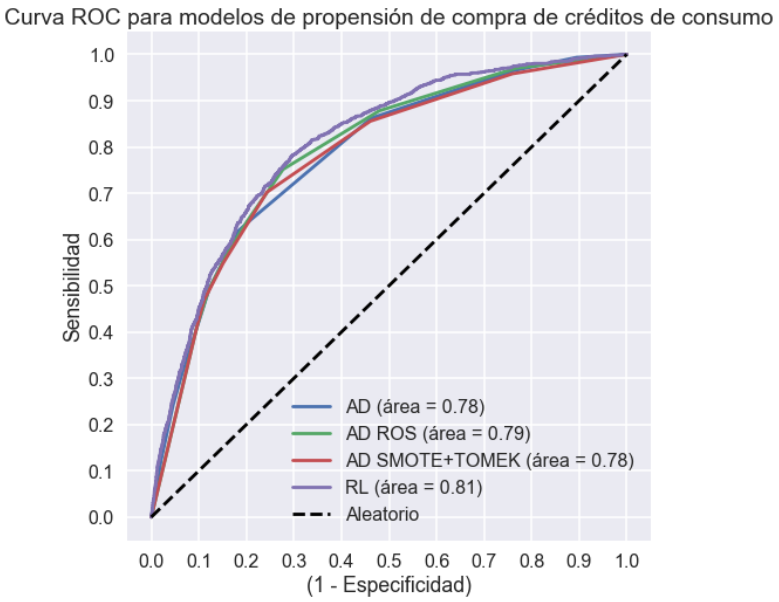
Los demás árboles de decisión se pueden observar en la **Ilustración 52** e **Ilustración 53**, ubicadas en anexos.

Regresión logística

Pasando al modelo de regresión logística, este se desempeña relativamente mejor en todas las métricas de desempeño, sin embargo, la sensibilidad de la clase negativa y la exactitud disminuye en 1 punto porcentual con respecto al árbol de decisión sin balanceo de datos como se puede observar en la Tabla 13.

La Ilustración 28 deja en evidencia como el modelo de regresión logística se desempeña levemente mejor a lo largo de toda la curva ROC, por lo tanto, se puede decir que la regresión logística se desempeña mejor en cuanto a la sensibilidad de la clase positiva en todas las probabilidades de corte.

Ilustración 28: Curva ROC para modelos de propensión de compra de créditos de consumo con su respectivo AUC (AD: Árbol de Decisión; RL: Regresión Logística).



Fuente: Elaboración propia.

Finalmente, se selecciona el modelo de regresión logística, porque, se puede seleccionar la cantidad de clientes con exactitud a través de la probabilidad de corte (lo cual es una ventaja al resolver la optimización propuesta más adelante) y se desempeña mejor a lo largo de toda la curva ROC.

Aunque parece que la diferencia es mínima, si se selecciona al 30% de los mejores clientes, con el modelo de árbol de decisión con Random Over Sampling se captura un 76,6% de la venta y con la regresión logística se captura un 78,5% de la venta. Esto equivale a una diferencia de 16 créditos de consumo o MM$188, es decir, un día completo de ventas.

Los coeficientes y significancias resultantes se encuentran en la Tabla 14 y en negrita se destacan las 15 variables con mayor coeficiente en valor absoluto, por lo tanto, son las variables que afectan más a la probabilidad de tomar un crédito de consumo, ya sea positivamente o negativamente.

Tabla 14: Modelo de propensión de créditos de consumo. Coeficientes, error estándar, z (estadístico t), p-valor e intervalo de confianza. En negrita, 15 variables con mayor coeficiente en valor absoluto.

Variable	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Constante	-4.674	0.090	-51.743	0.000	-4.851	-4.497
Realiza abonos TEF en el último año	0.317	0.085	3.718	0.000	0.150	0.484
Posee crédito de consumo que vence en menos de 6 meses	**0.519**	**0.052**	**10.043**	**0.000**	**0.418**	**0.621**
Realiza giros en tarjeta de débito en los últ. 6 meses	0.187	0.042	4.427	0.000	0.104	0.270
Porcentaje de pagos con cuotas en el último año = 0%	-0.190	0.052	-3.673	0.000	-0.291	-0.089
Porcentaje de pagos con cuotas en el último año = (0%, 20%]	0.099	0.048	2.043	0.041	0.004	0.194
Porcentaje de pagos con cuotas en el último año = (20%, 100%]	0.164	0.062	2.635	0.008	0.042	0.285
Inactivo en tarjeta de crédito	*	*	*	*	*	*
Ratio del promedio de la deuda de consumo en SBIF de los últ. 3 meses sobre los últ. 6 meses = (80%, 120%]	0.154	0.037	4.190	0.000	0.082	0.226
Sin deuda de consumo en SBIF, [0%,80%] y (120%, inf)	*	*	*	*	*	*
Días desde la última simulación de crédito de consumo = (0, 90]	**0.990**	**0.051**	**19.484**	**0.000**	**0.890**	**1.089**
Días desde la última simulación de crédito de consumo = (90, 366]	0.395	0.049	8.006	0.000	0.298	0.492
No ha realizado simulaciones de crédito de consumo en el último año	*	*	*	*	*	*
Porcentaje de cargos TEF a la misma persona en los últ. 12 meses = [0%, 30%]	0.195	0.064	3.069	0.002	0.071	0.320
Porcentaje de cargos TEF a la misma persona en los últ. 12 meses = (30%, 60%]	0.392	0.073	5.378	0.000	0.249	0.535
Porcentaje de cargos TEF a la misma persona en los últ. 12 meses = (60%, 90%]	**0.455**	**0.084**	**5.426**	**0.000**	**0.291**	**0.620**
Porcentaje de cargos TEF a la misma persona en los últ. 12 meses = (90%, 100%]	0.261	0.078	3.335	0.001	0.108	0.414
Inactivo en cargos TEF	*	*	*	*	*	*
Realiza apertura de consumo en los últ. 12 meses	**0.403**	**0.048**	**8.364**	**0.000**	**0.309**	**0.498**
Saldo vista mensual promedio de los últ. 12 meses = (M$700, MM$2]	-0.276	0.041	-6.721	0.000	-0.356	-0.195
Saldo vista mensual promedio de los últ. 12 meses = (MM$2, MM$5]	**-0.407**	**0.049**	**-8.354**	**0.000**	**-0.502**	**-0.311**
Saldo vista mensual promedio de los últ. 12 meses = (MM$5, MM$8]	**-0.495**	**0.080**	**-6.186**	**0.000**	**-0.652**	**-0.338**
Saldo vista mensual promedio de los últ. 12 meses = (MM$8, inf]	**-0.624**	**0.094**	**-6.615**	**0.000**	**-0.809**	**-0.439**
Saldo vista mensual promedio de los últ. 12 meses = [0, M$700]	*	*	*	*	*	*

Variable	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Promedio de la deuda de consumo en Banco de los últ. 12 meses = ($0, M$200]	0.476	0.085	5.631	0.000	0.310	0.642
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (M$200, M$400]	0.821	0.103	7.962	0.000	0.619	1.023
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (M$400, M$800]	1.020	0.097	10.557	0.000	0.830	1.209
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (M$800, MM$1.7]	1.109	0.093	11.877	0.000	0.926	1.292
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (MM$1.7, MM$3]	1.308	0.095	13.811	0.000	1.123	1.494
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (MM$3, MM$6]	1.492	0.092	16.262	0.000	1.312	1.671
Promedio de la deuda de consumo en Banco de los últ. 12 meses = (MM$6, inf]	1.786	0.089	20.103	0.000	1.612	1.960
Promedio de la deuda de consumo en Banco de los últ. 12 meses = 0$	*	*	*	*	*	*
Monto de inversiones comprobado del últ. Mes = (MM$3, inf]	-0.322	0.039	-8.312	0.000	-0.397	-0.246
Monto de inversiones comprobado del últ. Mes = [MM$0, MM$3]	*	*	*	*	*	*

Fuente: Elaboración propia.

Al igual que en el árbol de decisión con Random Over Sampling, nuevamente se repite como una de las variables más importantes la apertura de créditos de consumo en el último año, el promedio de la deuda de consumo en el Banco del último año, si ha realizado simulaciones de créditos de consumo y el saldo vista mensual promedio del último año.

Entre las 15 variables con mayor coeficiente en valor absoluto también se hace presente si el cliente posee un crédito de consumo que vence en menos de 6 meses y el porcentaje de cargos en transferencias electrónicas a la misma persona.

Nota: Algunos rangos de variables no existen, debido a que, estos no eran significativos al 95% de confianza y fueron eliminados, es decir, todas las variables finalmente seleccionadas son estadísticamente significativas al 95% de confianza.

Analizando los coeficientes,

1. Un cliente que ha realizado abonos TEF en los últimos 12 meses tiene mayor probabilidad de tomar un crédito de consumo que un cliente que no ha realizado abonos. Esta variable identifica a los clientes que utilizan su cuenta corriente, por lo tanto, un cliente que no ocupa su cuenta corriente con el Banco difícilmente va a tomar un crédito de consumo con él.

2. Un cliente que está realizando el pago de las últimas 6 cuotas o menos tiene una mayor probabilidad de tomar un crédito de consumo que un cliente que no se encuentre en esta situación. Esta variable identifica a los clientes que se han endeudado con el Banco y que tienen la oportunidad de volver a endeudarse.

3. Un cliente que ha realizado giros en su tarjeta de débito en los últimos 6 meses tiene mayor probabilidad de tomar un crédito de consumo que un cliente que no lo ha hecho.

4. Un cliente que ha pagado en cuotas alguna(s) de sus compras con tarjeta de crédito tiene mayor probabilidad de tomar un crédito con respecto a los que no usan la tarjeta de crédito, sin embargo, un cliente que nunca ha realizado compras sin cuotas en el último año disminuye su probabilidad. Por lo tanto, los clientes que están dispuestos a endeudarse en su tarjeta de crédito, también lo están para endeudarse en un crédito de consumo.

5. Un cliente que ha mantenido su deuda de consumo durante los últimos 12 meses tiene mayor probabilidad de tomar un crédito de consumo que un cliente que la ha disminuido, aumentado o no ha tenido deuda de consumo en los últimos 12 meses.

6. Un cliente que ha realizado una simulación de crédito de consumo tiene mayor probabilidad de tomar un crédito de consumo que un cliente que no ha realizado una en los últimos 12 meses, también se puede decir que un cliente que la realizó en los últimos 3 meses tiene más probabilidad que uno que la realizó hace más de 3 meses.

7. Un cliente que posee cargos en transferencias electrónicas (TEF) tiene una mayor probabilidad de tomar un crédito de consumo que un cliente que no posee cargos en los últimos 12 meses. Por otro lado, mientras mayor sea el porcentaje de cargos en transferencias electrónicas a su nombre aumenta la probabilidad de tomar un crédito de consumo, sin embargo, más allá del 90% de cargos a su nombre disminuye el aumento de la probabilidad. Esto puede deberse a que los clientes que se ven obligados a mover dinero entre sus cuentas probablemente necesiten un crédito de consumo, sin embargo, los clientes con más de un 90% de los cargos utilizan otro banco y si necesitan un crédito lo tomarían en el otro.

8. Un cliente que ha realizado una o más aperturas de consumo en el Banco en los últimos 12 meses, posee mayor probabilidad de tomar otro crédito de consumo que un cliente que no.

9. Un cliente que posee un saldo vista promedio mensual menor a $700.000 posee una probabilidad mayor a tomar un crédito de consumo que un cliente que posee un mayor saldo vista promedio mensual. Por otro lado, la probabilidad de tomar un crédito de consumo disminuye a medida que aumenta el saldo vista promedio mensual, debido a que, estos clientes no tienen la necesidad de un crédito de consumo.

10. Un cliente que posee una deuda de consumo con el Banco posee una probabilidad mayor a tomar un crédito de consumo que un cliente que no tiene deudas de consumo con el Banco. Por otro lado, la probabilidad de tomar un crédito de consumo aumenta a medida que aumenta la deuda de consumo con el Banco.

11. Un cliente que posee inversiones menores a $3.000.000 posee una mayor probabilidad de tomar un crédito de consumo que un cliente con mayor monto de inversiones comprobadas.

Por lo tanto, se puede identificar que el tipo de clientes que toma los créditos de consumo pre aprobados, en general son clientes que tienen un vínculo fuerte con el Banco, ya se han endeudado en el Banco, utilizan los distintos productos del Banco, no tienen demasiado dinero disponible en su cuenta y no realizan inversiones importantes.

6.6.1.2 Modelo de propensión de compra de tarjetas de créditos pre aprobadas Arboles de decisión

Los mejores resultados de los árboles de decisión con respecto al AUC obtenidos en la muestra de prueba, se entrenaron con los parámetros observados en la Tabla 15.

Tabla 15: Parámetros de parada para cada uno de los árboles de decisión. Modelo de propensión de tarjetas de crédito.

	Árbol de decisión		
Balanceo	**-**	**ROS**	**SMOTE + Tomek**
Máxima profundidad	5	6	6
Mínimo de observaciones en nodo padre	10%	-	-
Mínimo de observaciones en nodo hijo	8%	-	-
Reducción mínima de impureza	-	0,1%	0,1%

Fuente: Elaboración propia.

Los resultados de las métricas de desempeño para cada uno de los modelos de propensión a la toma de tarjetas de crédito se encuentran en la Tabla 16.

Tabla 16: Métricas de evaluación de árbol de decisión para los distintos métodos de balanceo de clases y regresión logística. Modelo de propensión de tarjetas de crédito.

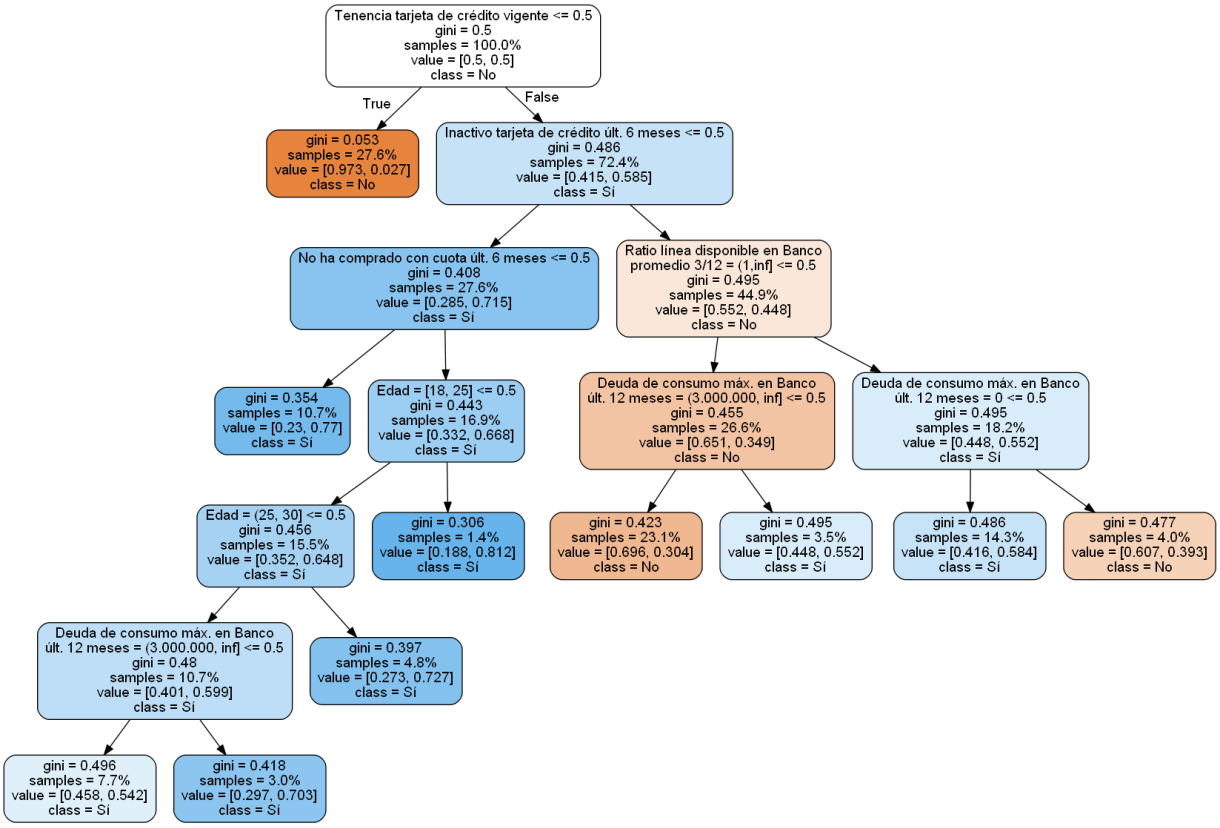
Árbol de decisión				Regresión logística
Balanceo	**-**	**ROS**	**SMOTE + Tomek**	**-**
Exactitud	0,76	0,77	0,77	**0,85**
NPV	**0,96**	**0,96**	**0,96**	0,95
Precisión	0,17	0,18	0,18	**0,22**
Especificidad	0,77	0,78	0,78	**0,88**
Sensibilidad	0,59	**0,63**	**0,63**	0,46
F1-Score	0,26	0,28	0,28	**0,30**
AUC	0,77	0,78	0,77	**0,80**

Fuente: Elaboración propia.

Para los árboles de decisión, mediante el Random Over Sampling se obtienen los mejores resultados con respecto a el área bajo la curva de ganancia. Se logra un aumento de 1 punto porcentual sobre el árbol de decisión sin balanceo de datos y con SMOTE + TOMEK.

Como se puede observar en la Ilustración 29, las variables que fueron seleccionadas por el árbol de decisión (con Random Over Sampling), corresponden a si el cliente posee una tarjeta de crédito vigente, si en los últimos 6 meses el cliente ha estado inactivo en la tarjeta de crédito, si ha comprado con cuotas en los últimos 6 meses, la edad del cliente, la deuda de consumo máxima en el Banco en los últimos 12 meses y el ratio de la línea disponible en Banco de los últimos 3 meses sobre los últimos 12 meses.

Ilustración 29: Árbol de decisión de modelo de propensión de compra de tarjetas de crédito con método Random Over Sampling.



Fuente: Elaboración propia.

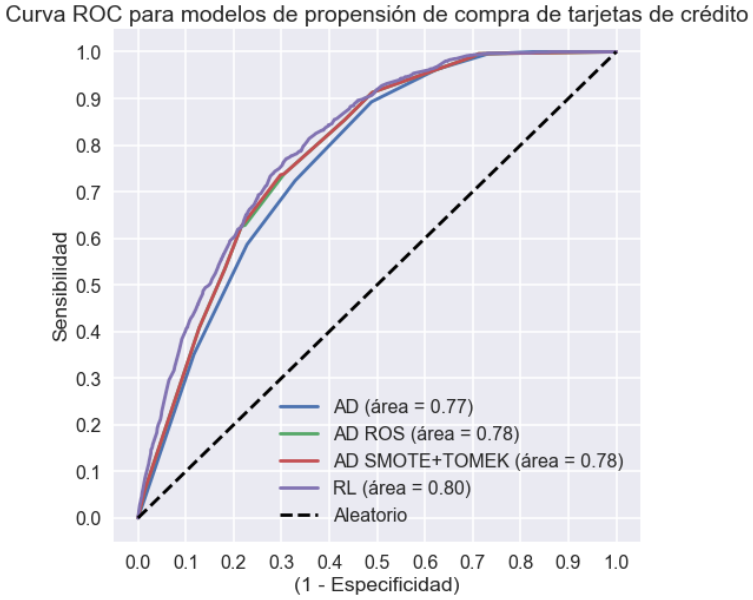
Los demás árboles de decisión se pueden observar en la **Ilustración 54** e **Ilustración 55**, ubicadas en anexos.

Regresión logística

Pasando al modelo de regresión logística, este no se desempeña mejor en todas las métricas de desempeño, ya que, empeora el valor predictivo negativo y la sensibilidad, sin embargo, en AUC se logra un aumento de 2 puntos porcentuales, en exactitud de 8 puntos porcentuales, en precisión y especificidad de 4 puntos porcentuales y en F1-Score de 2 puntos porcentuales con respecto al árbol de decisión con Random Over Sampling.

En la Ilustración 30, se puede observar como el modelo de regresión logística se desempeña levemente mejor a lo largo de toda la curva, por lo tanto, se puede decir que la regresión logística se desempeña mejor en cuanto a la sensibilidad de la clase positiva en todas las probabilidades de corte.

Ilustración 30: Curva ROC para modelos de propensión de compra de tarjetas de crédito con su respectivo AUC (AD: Árbol de Decisión; RL: Regresión Logística).



Fuente: Elaboración propia.

Si bien se observa que la regresión logística no tiene el mejor desempeño en cuanto a la sensibilidad de la clase positiva, que es la métrica más importante, gracias a que la predicción en este tipo de modelos es un valor continuo se puede seleccionar una probabilidad de corte que iguale a los demás modelos. Los resultados al modificar la probabilidad de corte son los observados en la Tabla 17.

Tabla 17: Resultados de métricas de desempeño, luego de modificar la probabilidad de corte en la regresión logística con el objetivo de igualar los resultados del árbol de decisión con Random Over Sampling.

	Árbol de decisión	Regresión logística
Balanceo	**ROS**	**-**
Exactitud	0,77	0,77
NPV	0,96	0,96
Precisión	0,18	0,18
Especificidad	0,78	0,78
Sensibilidad	0,63	0,63
F1-Score	0,28	0,28
AUC	0,78	0,80

Fuente: Elaboración propia.

Por lo tanto, se observa que la regresión logística puede obtener los mismos resultados que el árbol de decisión y permite seleccionar con exactitud la cantidad de clientes que se quiere priorizar a través de la probabilidad de corte, lo que facilita la optimización que se realiza posteriormente. Por lo que finalmente se selecciona la regresión logística.

En la Tabla 18 se pueden observar los coeficientes y significancia de las variables. En negrita se encuentran destacadas las 5 variables con mayor coeficiente en valor absoluto, por lo tanto, son las variables que afectan más a la probabilidad de tomar un aumento de cupo u obtención de tarjeta de crédito, ya sea positiva o negativamente.

Tabla 18: Modelo de propensión de tarjeta de crédito. Coeficientes, error estándar, z (estadístico t), p-valor e intervalo de confianza. En negrita, 5 variables con mayor coeficiente en valor absoluto.

Variable	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Constante	-8.038	0.611	-13.165	0.000	-9.235	-6.842
Género masculino	0.346	0.037	9.256	0.000	0.273	0.420
Saldo vista mensual promedio mayor a 0 en los últimos 12 meses	**1.215**	**0.209**	**5.821**	**0.000**	**0.806**	**1.624**
Tenencia de tarjeta de crédito vigente	**2.376**	**0.188**	**12.610**	**0.000**	**2.007**	**2.746**
Ratio línea de crédito disponible promedio de los últ. 3 meses sobre los últ 12 meses = [0%, 100%]	**1.701**	**0.608**	**2.795**	**0.005**	**0.508**	**2.893**
Ratio línea de crédito disponible promedio de los últ. 3 meses sobre los últ 12 meses = (100%, inf]	**2.042**	**0.609**	**3.353**	**0.001**	**0.849**	**3.236**
Sin línea de crédito disponible en los últ. 12 meses	*	*	*	*	*	*
Deuda de consumo máxima en Banco de los últ. 12 meses = ($0, M$300]	0.598	0.076	7.902	0.000	0.450	0.746
Deuda de consumo máxima en Banco de los últ. 12 meses = (M$300, MM$1]	0.751	0.082	9.113	0.000	0.589	0.912
Deuda de consumo máxima en Banco de los últ. 12 meses = (MM$1, MM$3]	0.892	0.083	10.731	0.000	0.729	1.055
Deuda de consumo máxima en Banco de los últ. 12 meses = (MM$3, inf]	**1.317**	**0.078**	**16.973**	**0.000**	**1.165**	**1.469**
Deuda de consumo máxima en Banco de los últ. 12 meses = $0	*	*	*	*	*	*

Variable	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Realiza al menos una compra con cuotas en los últ. 6 meses	1.037	0.048	21.746	0.000	0.943	1.130
No realiza compras con cuotas en los últ. 6 meses	0.553	0.044	12.423	0.000	0.466	0.640
Inactivo en tarjeta de crédito	*	*	*	*	*	*
Edad = (25, 30]	-0.607	0.066	-9.210	0.000	-0.736	-0.478
Edad = (30, 50]	-1.187	0.066	-18.021	0.000	-1.316	-1.058
Edad = (50, inf]	-1.485	0.077	-19.417	0.000	-1.635	-1.335
Edad = [18, 25]	*	*	*	*	*	*

Fuente: Elaboración propia.

Al igual que en el árbol de decisión con Random Over Sampling, nuevamente se repite como una de las variables más importantes la tenencia de tarjeta de crédito vigente y el ratio entre el promedio de la línea de crédito disponible en el Banco de los últimos 3 meses sobre los últimos 12 meses, sin embargo, la edad y la deuda de consumo máxima en el banco de los últimos 12 meses no se encuentra entre las 5 variables más importantes.

Entre las 5 variables con mayor coeficiente en valor absoluto ahora se hace presente el saldo vista mensual promedio de los últimos 12 meses.

Analizando los coeficientes,

1. Un hombre es más propenso en tomar un aumento de crédito de consumo que una mujer.

2. Tener un saldo vista mensual promedio mayor a 0, aumenta la probabilidad de tomar la oferta de aumento de cupo. Por lo tanto, los clientes que no utilizan su cuenta corriente difícilmente decidan aumentar su cupo en tarjeta de crédito.

3. Tener una tarjeta de crédito vigente es el factor que más aumenta la probabilidad de tomar la oferta de aumento de cupo, ya que, los clientes que no tienen una tarjeta de crédito vigente al tomar la oferta realizan la apertura de una tarjeta de crédito, sin embargo, muy pocos clientes lo hacen.

4. Los clientes que han aumentado su línea de crédito disponible poseen mayor probabilidad de tomar un aumento de cupo que un cliente que la ha disminuido, sin embargo, ambos poseen una mayor probabilidad que los clientes que en los últimos 12 meses no han tenido línea de crédito en el Banco.

5. A medida que aumenta la deuda de consumo máxima que el cliente ha tenido con el Banco en los últimos 12 meses también aumenta la probabilidad de tomar un aumento de cupo.

6. Clientes que han pagado alguna de sus compras con tarjeta de crédito en cuotas tienen mayor probabilidad de tomar un aumento de cupo que los clientes que no han realizado ninguna compra con cuotas, sin embargo, ambos clientes poseen

una mayor probabilidad de tomar un aumento de cupo que los clientes que no han utilizado la tarjeta de crédito en los últimos 6 meses.

7. Clientes jóvenes poseen una mayor probabilidad de tomar un aumento de cupo y a medida que aumenta la edad disminuye la probabilidad de tomar un aumento de cupo.

Por lo tanto, el tipo de clientes que toma los aumentos de cupo u obtención de tarjetas de crédito pre aprobados, generalmente son hombres jóvenes que utilizan la tarjeta de crédito y cuenta corriente, poseen deudas con el Banco y que han aumentado su línea de crédito en el sistema financiero (puede ser con el Banco o con otra institución).

6.6.1.3 Validación de modelos de propensión

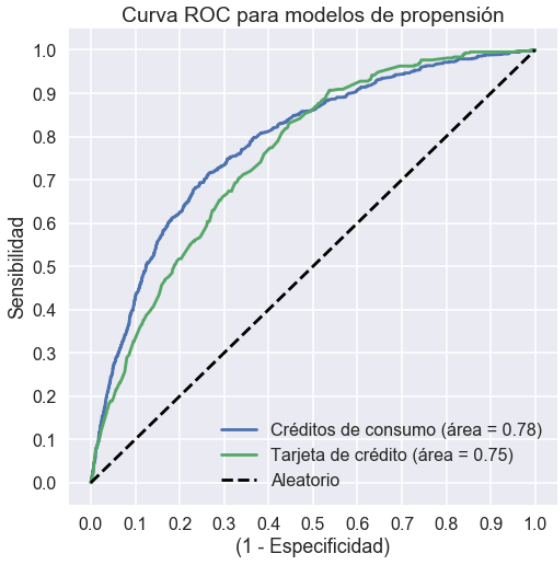
Para realizar la validación de los modelos, se utiliza la información de las ventas del tercer trimestre del año 2017 (julio, agosto y septiembre, ver Ilustración 24).

Es importante destacar que esta información no se toma en consideración para el ajuste de parámetros ni evaluación de los modelos, por lo tanto, sirve para verificar que no existe algun tipo de sobre ajuste.

Para calcular la probabilidad de que cada cliente compre un crédito de consumo o aumento de tarjeta de crédito se utiliza información desde julio de 2016 a junio de 2017.

En la Ilustración 31, se puede observar que los resultados no son lejanos a los obtenidos en la muestra de prueba, ya que, en la muestra de prueba se obtiene un AUC de 0,81 en para el modelo de propensión de crédito de consumo y un AUC de 0,80 para el modelo de propensión de tarjeta de crédito. Por lo tanto, se considera que es un buen resultado para la validación.

Ilustración 31: Curva de ROC para modelo de propensión de créditos de consumo y tarjeta de crédito en muestra de validación.



Fuente: Elaboración propia.

El área de inteligencia de negocios no tiene un modelo de propensión, sólo utiliza reglas básicas y el detalle de las prioridades se encuentra en la sección de Oportunidades y consecuencias.

Para poder comparar los resultados obtenidos a través de la priorización actual y los modelos de propensión, se toma en consideración que las primeras 4 prioridades corresponden a una predicción positiva, ya que, estas prioridades poseen las mejores tasas de respuesta. Luego se calculan los indicadores de los modelos al predecir positivamente el porcentaje de clientes que poseen estas prioridades, por ejemplo, si en las prioridades 1, 2, 3 y 4 se encuentra el 30% de los clientes, entonces se fija la probabilidad de corte de los modelos para predecir positivamente un 30%.

En la Tabla 2, se muestra la tasa de respuesta de créditos de consumo por prioridad asignada en la campaña del tercer trimestre del año 2017, se observa que se identifica al 28% de los clientes entre las cuatro primeras prioridades y que se captura el 47,3% de las ventas de créditos de consumo, esto equivale a que se tiene una precisión de 5,6%.

Al observar los indicadores del modelo de créditos de consumo para la muestra de validación en la Tabla 19, se destaca que con el 28% de los clientes se captura el 71% de la venta, por lo tanto, se considera un muy buen resultado, ya que, con el mismo porcentaje de clientes priorizados se captura un 23,7% más de la venta y se alcanza casi el doble de precisión.

Tabla 19: Métricas de evaluación de modelo de propensión de créditos de consumo en muestra de validación.

	Regresión logística
Exactitud	0,74
NPV	0,98
Precisión	0,10
Especificidad	0,74
Sensibilidad	0,71
F1-Score	0,18
Porcentaje de predicción positiva	0,28

Fuente: Elaboración propia.

La matriz de confusión para el modelo de propensión de créditos de consumo en la muestra de validación se puede observar en la Tabla 20.

Tabla 20: Matriz de confusión en porcentajes para modelo de propensión de créditos de consumo en muestra de validación.

		Actual	
		Positiva	**Negativa**
Predicción	**Positiva**	2,92%	25,08%
	Negativa	1,20%	70,80%

Fuente: Elaboración propia.

Al analizar los errores tipo 1 y tipo 2 se observa lo siguiente:

- El error de tipo 1 corresponde al 25,08% de los clientes, es decir, si se priorizara a los clientes bajo estas condiciones, un 25,08% de los clientes serían priorizados en vano y se estaría generando un esfuerzo en vano. Sin embargo, este error es el menos costoso y hoy en día es un costo hundido, ya que, se gestiona al 72% de los clientes en promedio.

- El error de tipo 2 corresponde al 1,20%, es decir, no se priorizaría al 1,20% de los clientes que luego tomarían un crédito de consumo, por lo tanto, este es el error más importante para este problema y se espera que este sea lo más bajo posible.

Para el caso de los aumentos de cupo en tarjeta de crédito, en la **Tabla 27** se muestran las tasas de respuesta por prioridad asignada en la campaña del tercer trimestre del año 2017, se observa que se identifica al 20,0% de los clientes entre las cuatro primeras prioridades y que se captura el 35,6% de las ventas de tarjetas de créditos, esto equivale a una precisión del 6,8%.

Al observar los indicadores del modelo de tarjetas de crédito para la muestra de validación en la Tabla 21, se destaca que con el 20,0% de los clientes se captura el 50% de la venta, por lo tanto, también se considera que es un buen resultado, ya que, con el mismo porcentaje de clientes priorizados se captura un 14,4% más de la venta y se mejora la precisión en un 3,2%.

Tabla 21: Métricas de evaluación de modelo de propensión de tarjetas de crédito en muestra de validación.

	Regresión logística
Exactitud	0,80
NPV	0,98
Precisión	0.10
Especificidad	0,81
Sensibilidad	0,50
F1-Score	0,16
Porcentaje de predicción positiva	0,20

Fuente: Elaboración propia.

La matriz de confusión para el modelo de propensión de tarjetas de crédito en la muestra de validación se puede observar en la Tabla 22.

Tabla 22: Matriz de confusión en porcentajes para modelo de propensión de tarjetas de crédito en muestra de validación.

		Actual	
		Positiva	**Negativa**
Predicción	**Positiva**	1,95%	18,06%
	Negativa	1,95%	78,04%

Fuente: Elaboración propia.

Al analizar los errores tipo 1 y tipo 2 se observa lo siguiente:

- El error de tipo 1 corresponde al 18,06% de los clientes, es decir, si se priorizara a los clientes bajo estas condiciones, un 18,06% de los clientes serían priorizados en vano y se estaría generando un esfuerzo en vano. Sin embargo, este error no es el menos costoso y hoy en día es un costo hundido, ya que, el 72% de los clientes se gestiona en promedio.

- El error de tipo 2 corresponde al 1,95%, es decir, no se priorizaría al 1,95% de los clientes que luego tomarían un aumento de cupo en tarjeta de crédito, por lo tanto, este es el error más importante para este problema y se espera que este sea lo más bajo posible.

Es importante destacar, que el modelo de propensión a la compra de créditos de consumo se desempeña levemente mejor, ya que, el error de tipo 2 que es el más caro es menor y se obtiene una mejora de 23,7% en captura de la venta de créditos de consumo mientras que el modelo de tarjetas de crédito sólo obtiene una mejora de 14,4%.

Considerando que los modelos se desempeñan mejor que la priorización actual en los grupos de mayor prioridad y se obtienen los resultados esperados en los errores de tipo 2, ya que, estos no alcanzan un 2% probabilidad, los resultados de los modelos son satisfactorios.

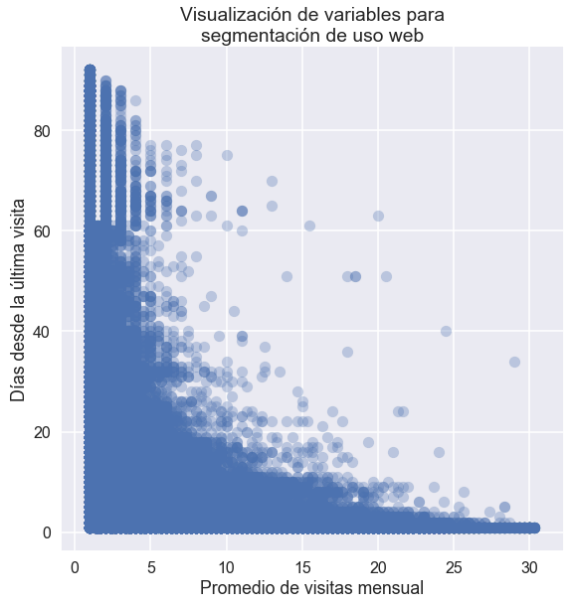
6.6.2 Modelos de segmentación

6.6.2.1 Segmentación de clientes pre aprobados de acuerdo con uso del sitio web

Para realizar esta segmentación se observa el comportamiento de los clientes pre aprobados de los últimos 3 meses antes de las campañas. Las variables analizadas y que serán las variables de entrada del algoritmo K-Means corresponden al promedio de días mensual en que visita el sitio web y la cantidad de días desde la última visita.

En el gráfico de la Ilustración 32, se observa que existen clientes que poseen un promedio de visitas mensual medio, sin embargo, no visitan el sitio web hace más de un mes, es por esta razón, que es importante tomar en consideración estas dos variables.

Ilustración 32: Visualización de variables utilizadas en segmentación de uso web por clientes pre aprobados.

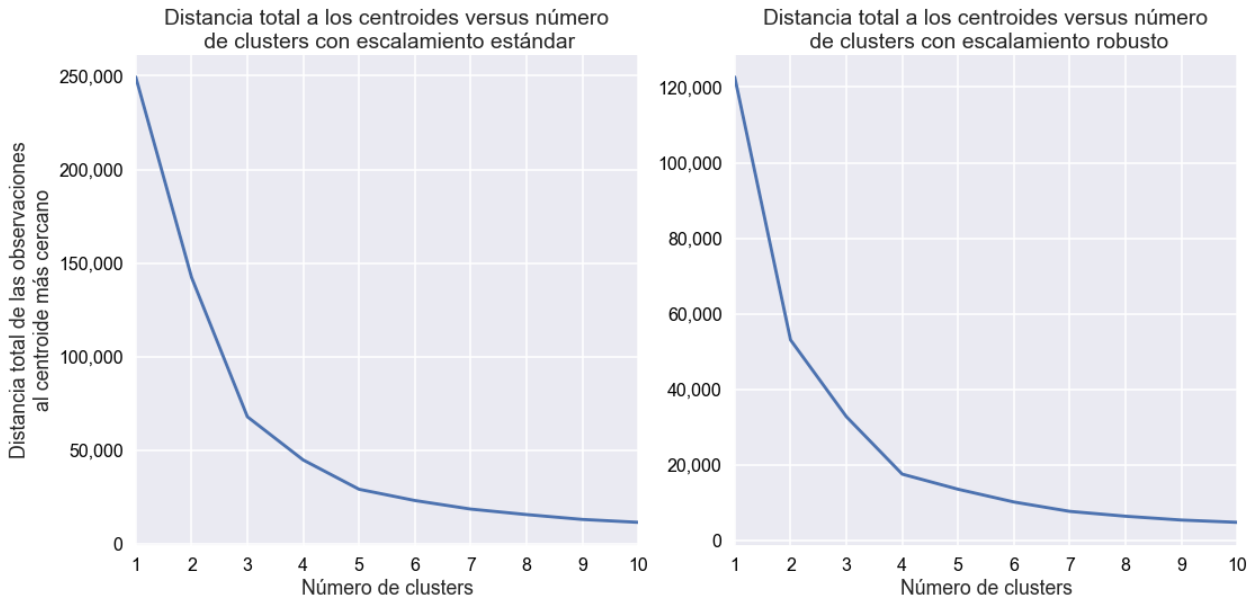


Fuente: Elaboración propia.

Para determinar el número de segmentos se utiliza el método del codo. En la figura de la izquierda en la Ilustración 33, se observa que con estandarización de los datos y 5 segmentos la suma de las distancias se vuelve significativamente menor y al agregar

segmentos este valor no decrece significativamente. Análogamente se observa que con 4 segmentos la distancia total se minimiza para el caso del escalamiento robusto.

Ilustración 33: Suma total de la distancia entre las observaciones y el centroide más cercano en función del número de clusters para los dos tipos de escalamiento para segmentación de uso del sitio web.

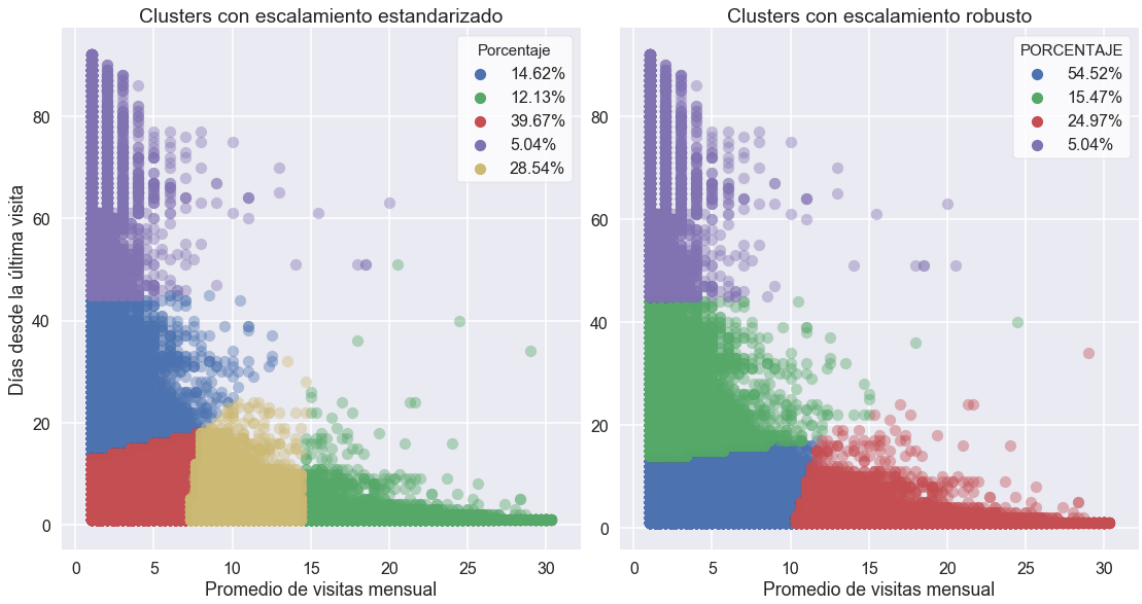


Fuente: Elaboración propia.

Se conoce que el escalamiento de los datos impacta en los resultados obtenidos por el algoritmo K-Means. Como se puede observar en la Ilustración 34 los resultados de la segmentación con escalamiento estandarizado[2] muestra que existen observaciones que poseen un alto promedio de visitas mensual, pero que no han visitado la página web hace más de 30 días a pesar de que son segmentados en el grupo de uso habitual de la página web. Esto se logra corregir con el escalamiento robusto de los datos[3], el cual entrega los resultados del segundo gráfico de la Ilustración 34.

[2] Se resta el promedio y se divide por la desviación estándar.
[3] Se resta la mediana y se divide por el rango inter-cuartil.

Ilustración 34: K-Means con escalamiento estandarizado y robusto de variables para segmentación de uso del sitio web.



Fuente: Elaboración propia.

Utilizando finalmente la segmentación con escalamiento robusto, es posible identificar cuatro grupos de clientes que se encuentran activos en los últimos 3 meses y que representan al 85% de los clientes pre aprobados, ya que, un 15% de los clientes no han visitado el sitio web en los últimos 3 meses antes de las campañas.

Del 85% de los clientes activos en los últimos 3 meses antes de las campañas, un 25% de los clientes pre aprobados hacen uso del sitio web habitualmente, un 55% hace un uso normal, 16% hace un uso bajo y un 5% hace un uso muy bajo del sitio web.

Tabla 23: Centroides de segmentación de uso del sitio web con algoritmo de K-Means y escalamiento robusto.

Segmento	Promedio de visitas mensual	Días desde la última visita	Porcentaje de clientes
Rojo Uso alto	15,47	1,80	24,97%
Azul Uso medio	5,62	4,09	54,52%
Verde Uso bajo	2,46	23,66	15,47%
Morado Uso muy bajo	1,56	63,69	5,04%

Fuente: Elaboración propia.

Como se puede observar en la Tabla 23, el promedio de visitas mensual de los clientes más habituales (uso alto) es de 15 días al mes y se puede decir que este tipo de clientes visita el sitio web aproximadamente 4 veces a la semana, lo que es confirmado

y se relaciona con que el promedio de los días desde la última visita de este grupo corresponda a dos días.

Los clientes normales (uso medio) en general visitan el sitio web una o dos veces a la semana, lo que también es respaldado por el promedio de los días desde la última visita.

Los clientes de uso bajo utilizan el sitio web una o dos veces cada dos semanas, sin embargo, el promedio de días desde la última visita da a entender que estos clientes en general no han visitado el sitio hace más de 3 semanas, por lo tanto, a partir de este segmento se observa que los clientes comienzan a alejarse de la utilización del sitio web.

Por último, el grupo de clientes de uso muy bajo, deberían visitar el sitio web una vez cada dos semanas, sin embargo, los días desde la última visita indican que en general estos clientes no han visitado el sitio web hace más de dos meses, por lo que, este tipo de clientes es muy probable que ya no utilice el sitio web o lo deje de utilizar.

Por lo tanto, es importante hacer llegar a los clientes del grupo de uso muy bajo (segmento Morado) y clientes inactivos la promoción a través de otros medios, debido a que, es posible que la oferta no la vean en más de dos meses, siendo que las campañas sólo duran 3 meses.

6.6.2.2 Segmentación de clientes pre aprobados de acuerdo con la apertura de correos electrónicos

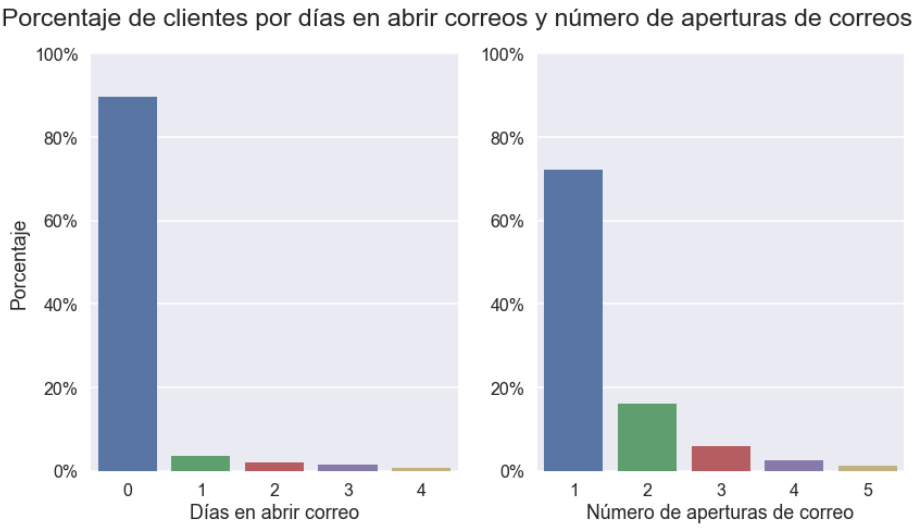
Las variables que se encuentran disponibles para analizar el uso de correo electrónico son las siguientes:

1. Cantidad de veces que el cliente abre el correo electrónico
2. Fecha del envió
3. Fecha de la primera apertura

También se construye una variable que cuenta los días que se demora el cliente en abrir el correo electrónico.

Al observar la cantidad de aperturas que realizan los clientes en la Ilustración 35, se observa que no existe una variabilidad considerable, ya que, de los clientes que abren los correos electrónicos, el 88% de los clientes realiza 1 o 2 aperturas del correo. Tampoco se observa una variabilidad en los días que se demoran en abrir los correos electrónicos, incluso un 90% realiza la apertura el mismo día en que se envía el correo. Por lo tanto, estas variables no se utilizarán para hacer el perfilamiento.

Ilustración 35: Porcentaje de clientes por días en abrir correos electrónicos y número de aperturas de correos electrónicos.

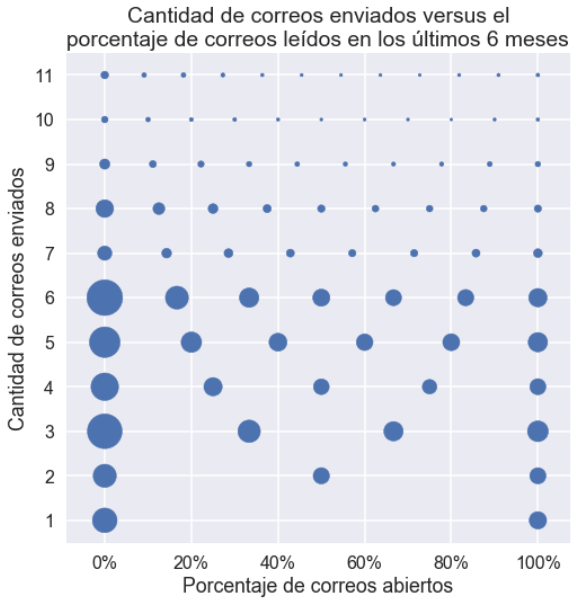


Fuente: Elaboración propia.

Por otro lado, la cantidad de correos enviados a los clientes en los últimos 6 meses varía para cada cliente, es por esta razón que un cliente que lea el 100% de los correos, pero que se le ha enviado sólo un correo, no puede ser considerado de igual forma que un cliente que lea el 100% de los 6 correos que se le han enviado.

Está situación queda en evidencia en la Ilustración 36, donde el tamaño de cada punto indica la frecuencia de clientes, es decir, a la mayoría de los clientes pre aprobados en los últimos 6 meses, se les ha enviado 6 correos y han leído el 0% de estos.

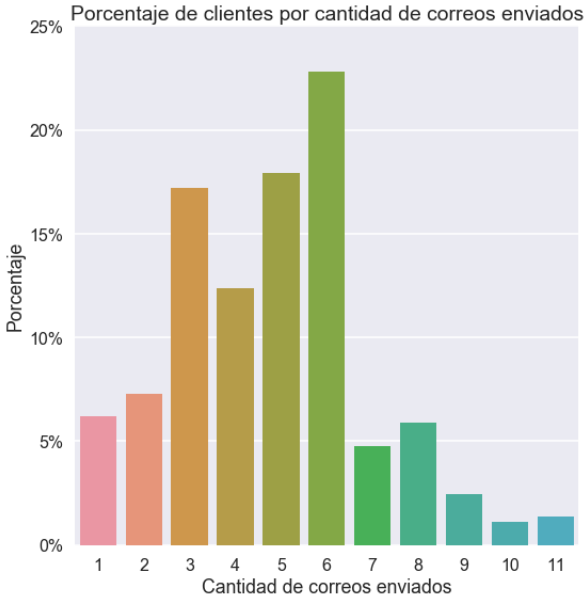
Ilustración 36: Cantidad de correos enviados en función del porcentaje de correos leídos. El tamaño de cada punto representa la frecuencia.



Fuente: Elaboración propia.

Como se puede observar en la Ilustración 38 al 84% de los clientes pre aprobados se les ha enviado 6 o menos correos en los últimos 6 meses, por esta razón, se decide estudiar el comportamiento de los clientes con respecto a los últimos 6 correos enviados en los últimos 6 meses.

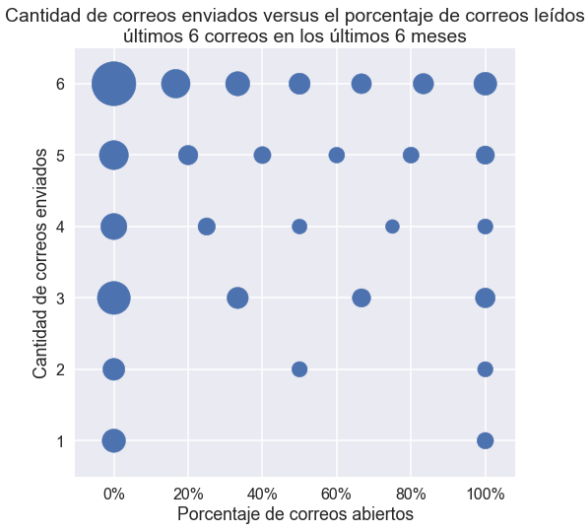
Ilustración 37: Porcentaje de clientes en función de la cantidad de correos enviados.



Fuente: Elaboración propia.

En la Ilustración 38, se puede observar el comportamiento de los clientes, con respecto a la cantidad de correos y el porcentaje de correos abiertos. También se tiene que la mayoría de los clientes no ha leído ninguno de los últimos 6 correos que se les ha enviado.

Ilustración 38: Cantidad de correos enviados en función del porcentaje de correos leídos de los últimos 6 correos en los últimos 6 meses. El tamaño de cada punto representa la frecuencia.

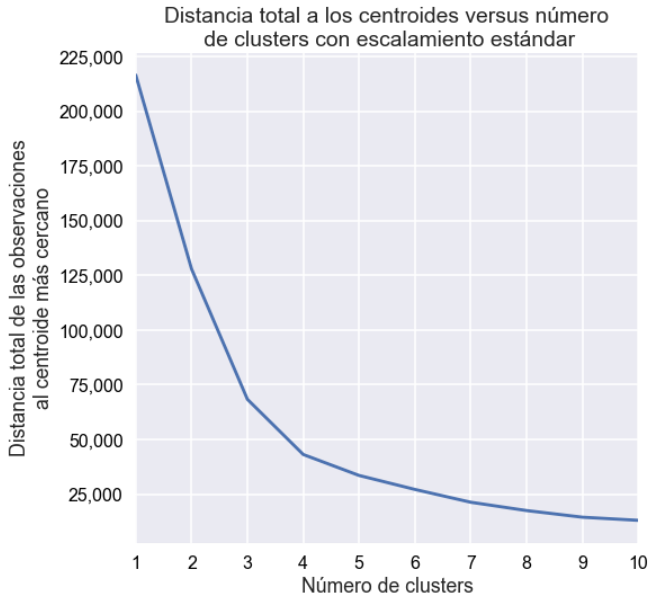


Fuente: Elaboración propia.

En este caso no fue necesario realizar un escalamiento de los datos robusto, debido a que, se obtienen los mismos resultados que con escalamiento estándar.

Para determinar el número de segmentos se utiliza el método del codo. En el gráfico de la Ilustración 39, se observa que con 4 segmentos la suma de las distancias se vuelve significativamente menor, sin embargo, este valor no disminuye de la misma forma con más clusters.

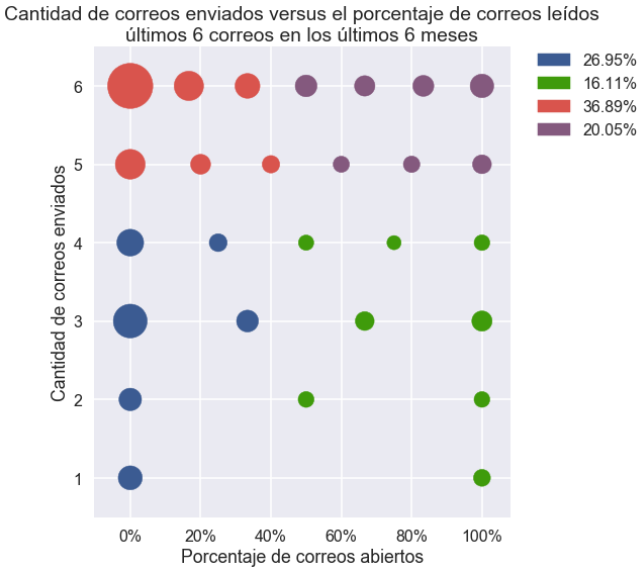
Ilustración 39: Suma total de la distancia entre las observaciones y el centroide más cercano en función del número de clusters en segmentación de uso de correo electrónico.



Fuente: Elaboración propia.

Los resultados de la segmentación con 4 grupos se observan en la Ilustración 40. Se observa que la mayoría de los clientes, al 36,89% se les ha enviado 5 o 6 correos, sin embargo, abren menos de un 50% de los correos. Por otro lado, al 20,05% de los clientes que se les han enviado 5 o 6 correos lee el 50% o más de los correos. También se observa que el 43,06% de los clientes se les han enviado 4 o menos correos electrónicos.

Ilustración 40: Resultados de segmentación de uso de correo electrónico con 4 clusters mediante el método K-means.



Fuente: Elaboración propia.

En la Tabla 24, se observa que los clientes del segmento rojo y morado son los clientes que se les han enviado en promedio 6 correos, por lo tanto, el comportamiento de ellos se debiese mantener en el tiempo.

Tabla 24: Centroides de segmentación de uso de correo electrónico con algoritmo de K-Means y escalamiento estandarizado.

Segmento	Porcentaje de correos abiertos	Cantidad de correos enviados	Porcentaje de clientes
Azul Uso reciente bajo	6,3%	2,8	26,95%
Rojo Uso constante bajo	10,8%	5,7	36,89%
Morado Uso constante alto	78,1%	5,7	20,05%
Verde Uso reciente alto	81,6%	2,8	16,11%

Fuente: Elaboración propia.

El segmento de uso constante bajo, en promedio abre el 11% de los correos, por lo tanto, se considera que, al enviar un nuevo correo a este tipo de clientes, este no va a ser abierto.

El segmento de uso constante alto, en promedio lee el 78% de los correos, por lo tanto, a estos clientes se considera que los correos que se les envían van a ser leídos.

Finalmente, se observa que en promedio al segmento de uso reciente bajo y alto se les han enviado sólo 3 correos en 3 meses, es decir, menos de un correo mensual, por lo tanto, se considera que a estos clientes se les debería seguir enviando correos hasta haber enviado al menos 5 correos y observar si leen más o menos del 50% de los correos.

Al realizar la intersección de los clientes que son inactivos en el uso del sitio web o que hacen un uso muy bajo del sitio web (segmento Morado) con los clientes que son inactivos en el uso del correo electrónico o hacen un uso constantemente bajo del correo electrónico (segmento Rojo), estos corresponden al 14% de los clientes en promedio a lo largo de las campañas estudiadas, es decir, casi la quinta parte de los clientes podrían no recibir la oferta.

- En conclusión, se debe gestionar a los clientes que pertenecen a esta intersección a través de los ejecutivos con el objetivo de no perder una venta porque el cliente no recibe la oferta en los 3 meses que dura la campaña.

7 Optimización del grupo de clientes priorizados en la gestión de los ejecutivos

Con el objetivo de priorizar en la gestión a través de los ejecutivos a los clientes que potencialmente pueden cursar grandes montos en los productos pre aprobados, poseen una alta probabilidad de tomar los productos y tomando en consideración que los clientes que no son contactados a través de los ejecutivos tal vez no vean la oferta a través de los otros medios de contacto, se plantea un problema de optimización lineal entero.

El objetivo del problema es maximizar la utilidad esperada de los clientes priorizados y no priorizados, asignando el conjunto óptimo de clientes al grupo priorizado.

Los parámetros del problema son los siguientes:

- Se define p_{ij} como la probabilidad de que el cliente i tome el producto j. El valor de este parámetro se obtiene a través del modelo de propensión para cada producto.
- Se define o_j como el monto pre aprobado en el producto j.
- Se define TM_j como el transformador monetario del producto j. Este transformador monetario lleva el monto de cada producto a una misma base para poder ser comparados.
- Se define s_i como una variable binaria, que toma el valor de 1 si el cliente pertenece al segmento de uso de correo electrónico rojo (bajo porcentaje de apertura de correos electrónicos y alto número de correos enviados) o inactivo y al mismo tiempo pertenece al segmento de uso de sitio web morado (bajo promedio de visitas mensual y alto número de días desde la última visita) o inactivo. En caso contrario toma el valor de 0.
- Se define c como una penalización por no priorizar a los clientes, este parámetro toma un valor entre 0 inclusivo y 1 no inclusivo, por lo tanto, la utilidad esperada del cliente cuando no es priorizado es una parte de la utilidad esperada de cuando el cliente es priorizado. El caso en que c es igual a 1 no es óptimo, debido a que, priorizar a un cliente otorga la misma utilidad que no priorizarlo y la solución corresponde a priorizar sólo a los clientes con $s_i = 1$.

La variable de decisión corresponde a X_i que toma el valor de 1 si el cliente es priorizado o 0 en caso contrario.

La única restricción del problema corresponde al número k de clientes que se quiere priorizar, en este caso, este parámetro corresponde al 72% de los clientes, ya que,

como en promedio se gestiona al 72% de los clientes, se espera que estos clientes sean gestionados.

La formulación del problema se expresa de la siguiente forma:

$$Máx \quad \sum_{i=1}^{n}\left[\left(\sum_j p_{ij} * o_j * TM_j\right) * X_i + c * \left(\sum_j p_{ij} * o_j * TM_j\right) * (1 - s_i) * (1 - X_i)\right]$$

$$s.a. \quad \sum_{i=1}^{n} X_i = k$$

- El primer término de la función objetivo corresponde a la suma de la utilidad esperada de cada producto para cada cliente. Este término se activa cuando el cliente es priorizado y se vuelve 0 cuando el cliente no es priorizado.

- El segundo término, corresponde a la suma de la utilidad esperada de cada producto para cada cliente penalizada por el parámetro c. Este término se activa cuando el cliente no es priorizado y no pertenece a los segmentos de uso bajo o inactivo en correo electrónico y sitio web.

En la Tabla 25, se puede observar un ejemplo del resultado de optimizar a 10 clientes con un parámetro c igual a 0, 0,3 y 0,99 y un parámetro k igual a 6. Cabe destacar que los clientes se encuentran ordenados por utilidad esperada de cada cliente al priorizarlo, es decir, están ordenados por $\sum_j p_{ij} * o_j * TM_j$.

Tabla 25: Ejemplo de resultado de optimización cambiando el parámetro c.

Id	Prob. CC	Prob. TC	Oferta CC	Oferta TC	s_i	xi (c = 0)	xi (c = 0,3)	xi (c = 0,99)
1	80,5%	92,8%	$6.320.910	$6.897.733	0	1	1	1
2	74,4%	78,3%	$3.747.681	$9.076.744	0	1	1	1
3	46,3%	87,4%	$1.936.684	$5.442.317	0	1	1	1
4	96,9%	62,1%	$1.752.713	$4.995.236	0	1	1	0
5	84,0%	19,5%	$8.867.103	$1.352.593	0	1	0	0
6	73,9%	68,1%	$2.867.180	$3.468.279	0	1	0	0
7	72,2%	5,4%	$9.949.246	$5.193.140	0	0	0	0
8	72,1%	51,5%	$1.655.524	$4.859.473	1	0	1	1
9	25,0%	32,9%	$1.363.389	$6.688.028	1	0	1	1
10	74,3%	40,4%	$221.469	$5.566.920	1	0	0	1

Fuente: Elaboración propia.

- Para el caso extremo en que c es igual a 0, la función objetivo busca maximizar la utilidad esperada de los clientes sin considerar a los clientes que no son priorizados, por esta razón, la solución óptima prioriza a los 6 clientes con mayor utilidad esperada al priorizarlos.

- Para el caso en que c es igual a 0,3, la función objetivo si toma en cuenta a los clientes que no son priorizados. Se puede observar que no se priorizan a los clientes 5 y 6, debido a que, como estos clientes no pertenecen a los segmentos de bajo uso o inactivo en los canales digitales ($s_i \neq 1$) se prefiere priorizar a los clientes 8 y 9, que en el caso en que no son priorizados su utilidad esperada es 0, ya que, s_i es igual a 1 para estos clientes. Desde este punto de vista, el cliente 4 podría no priorizarse y priorizar al cliente 10, sin embargo, la utilidad esperada del cliente 10 es demasiado baja para priorizarlo, es decir, la suma de la utilidad esperada del cliente 4 al no priorizarlo y la utilidad esperada del cliente 10 al priorizarlo, es menor a la utilidad esperada de priorizar al cliente 4 y no priorizar al cliente 10.

- Para el otro caso extremo, en que c es igual a 0,99, la función objetivo prioriza a todos los clientes con s_i igual a 1 y luego prioriza a los clientes con mayor utilidad esperada al priorizarlos.

Por lo tanto, mientras menor sea el parámetro c menor importancia se les da a los clientes con s_i igual a 1 y mientras mayor sea el parámetro c mayor importancia se les da a estos clientes.

Para determinar este parámetro, se observa cómo afecta la gestión en la tasa de respuesta de cada producto a lo largo de todas las campañas con datos disponibles.

Para los clientes con oferta de tarjeta de crédito y que son gestionados se tiene una tasa de respuesta de 5,7% y los no gestionados de 2,3%, por lo tanto, se tiene que la tasa de respuesta de los clientes no gestionados es un 40% de la tasa de respuesta de los clientes gestionados.

Para la tarjeta de crédito se tiene que los clientes gestionados y con oferta de tarjeta de crédito poseen una tasa de respuesta de 9,3% y los no gestionados de 4,3%, por lo tanto, se tiene que la tasa de respuesta de los clientes no gestionados es un 46% de la tasa de respuesta de los no gestionados.

En promedio la tasa de respuesta de los clientes no gestionados corresponde al 43% de la tasa de respuesta de los clientes gestionados, por lo tanto, el parámetro c se fija en 0,43.

Un ejemplo del resultado de la optimización se muestra en la Tabla 26, donde se tiene la variable de decisión y la utilidad esperada de cada cliente. Se puede distinguir que los últimos clientes poseen una utilidad esperada de 0, debido a que, estos clientes pertenecen al grupo de clientes inactivo o de uso muy bajo de los canales digitales, por lo tanto, al no ser priorizados su utilidad esperada es 0.

Tabla 26: Ejemplo del resultado de la optimización.

Id Cliente	Priorizar	Utilidad esperada
52	1	$ 25,360,200
3243	1	$ 18,944,871
8640	1	$ 14,438,980
100	1	$ 4,704,374
5463	1	$ 3,575,174
582	1	$ 2,239,223
…	…	…
…	…	…
97	0	$ 248,992
1	0	0
813	0	0
15369	0	0

Fuente: Elaboración propia.

8 Diseño experimental

8.1 Factores y niveles

El factor que se desea estudiar corresponde a la priorización entregada a los ejecutivos en las campañas de promoción de créditos de consumo pre aprobados y aumentos de cupo de tarjetas de crédito pre aprobadas.

Los niveles de priorización que se desean estudiar corresponden a la priorización actual y la priorización planteada en la memoria.

8.2 Grupos de control y tratamiento

La promoción de los créditos de consumo pre aprobados y aumentos de cupo de tarjeta de crédito pre aprobado se realiza a través de e-mail, sitio web, SMS y ejecutivos.

Los canales por los que se comunica la promoción no corresponden a un factor, por lo tanto, estos se mantienen intactos y se realiza la promoción a través de todos los canales.

Los clientes tienen asignada una prioridad de gestión, por lo que los ejecutivos realizan un mayor esfuerzo por gestionar a los clientes con una mayor prioridad.

Los grupos con los que se cuenta son los siguientes:

- **Grupo de control:** No realizar la gestión y no contactar a los clientes a través de ningún medio de comunicación.
- **Grupo de tratamiento 1:** Realizar la gestión utilizando la priorización propuesta.
- **Grupo de tratamiento 2:** Realizar la gestión utilizando la priorización actual.

8.3 Hipótesis por validar

Hipótesis 1: “La priorización propuesta aumenta la tasa de respuesta de créditos de consumo en comparación con la priorización actual”

Hipótesis 2: “La priorización propuesta aumenta la tasa de respuesta de tarjetas de crédito en comparación con la priorización actual”

Para poder validar las hipótesis, se utiliza un test de proporciones, el cual fue detallado en el marco conceptual, en este contexto, la tasa de respuesta se refiere a el número de ventas del producto estudiado dividido por el número de clientes, por lo tanto, el número de ventas y número de clientes varía en cada grupo.

8.4 Muestra de clientes

Al reemplazar en la siguiente formula (ver marco teórico):

$$n = \frac{\chi^2 * P * (1 - P)}{d^2}$$

- χ^2 es igual a 1,96 para un nivel de confianza del 95%
- P igual a 0,5 para obtener el tamaño máximo de muestra
- d igual a 5%

Se obtiene que el tamaño necesario de la muestra es de 384 clientes. Entonces, tanto el grupo de control, como los de tratamiento, deben tener al menos 384 clientes.

Se plantea utilizar el 10% de los clientes como grupo de control, el 40% de los clientes como grupo de tratamiento 1 y el 40% como grupo de tratamiento 2, ya que, el número de clientes mínimo es respetado y se tiene la holgura suficiente en el grupo de control en caso de que el número de clientes se reduzca.

Para que los resultados no sean sesgados, las muestras de los distintos grupos de tratamiento y control deben tener las siguientes similitudes.

1. **Similar distribución de ofertas,** existen 7 combinaciones de ofertas de productos.
2. **Similar distribución de clientes por segmento,** en base al comportamiento transaccional de los clientes, existen 6 segmentos de clientes.
3. **Similar distribución de probabilidades de propensión** a créditos de consumo y tarjetas de crédito.

Por ejemplo, en el tercer trimestre del año 2017 el grupo de control y los grupos de tratamientos debiesen haber tenido las siguientes distribuciones:

1. 48,55% de clientes sólo con oferta de crédito de consumo, 0,05% sólo con oferta de línea de sobregiro, 4,29% sólo con oferta de tarjeta de crédito, 0,79% con oferta de crédito de consumo y línea de sobregiro, 45,23% con oferta de crédito de consumo y tarjeta de crédito, 0,03% con oferta de línea de sobregiro y tarjeta de crédito y 1,06% de clientes con oferta de los tres productos.

2. 18,28% pertenecen al segmento 1, 43,08% al segmento 2, 9,86% al segmento 3, 2,08% al segmento 4, 15,20% al segmento 5 y 11,51% al segmento 6.

3. La distribución de probabilidades para créditos de consumo y tarjeta de crédito se puede observar en la Ilustración 25.

8.5 Reporte a ejecutivos

Cada ejecutivo gestiona a cada uno de los clientes de su cartera, por lo tanto, existen dos formas de crear a los grupos de control y tratamientos.

La primera, es seleccionar 3 grupos de ejecutivos de tal forma que la cantidad de clientes coincida con un 10%, 40% y 40% de clientes para los grupos de control, tratamiento 1 y tratamiento 2 respectivamente. Sin embargo, esto puede afectar la validez del experimento, ya que, existen distintos tipos de ejecutivos que podrían sesgar los resultados.

La segunda opción y la que se recomienda utilizar, es seleccionar el 10%, 40% y 40% de los clientes, sin tomar en cuenta al ejecutivo, pero existe la posibilidad de que un ejecutivo posea clientes de los dos tipos de priorización, para solucionar este problema, se plantea mezclar las dos priorizaciones.

Para realizar la combinación, los clientes priorizados con la optimización se ordenan de acuerdo con su utilidad esperada y se asignan los clientes según la distribución de las 7 prioridades actuales, para entender esto observar el esquema de la Ilustración 41.

Ilustración 41: Esquema para combinar la priorización actual con la propuesta.

Id Cliente	Priorizar	Utilidad esperada	
52	1	$ 25,360,200	0,5% prioridad 1
3243	1	$ 18,944,871	3,0% prioridad 2
8640	1	$ 14,438,980	
100	1	$ 4,704,374	5,3% prioridad 3
5463	1	$ 3,575,174	
582	1	$ 2,239,223	18,5% prioridad 4
…	…	…	
…	…	…	38,8% prioridad 5
97	0	$ 248,992	
1	0	0	10,1% prioridad 6
813	0	0	
15369	0	0	23,7% prioridad 7

Fuente: Elaboración propia.

Finalmente, se deben ordenar aleatoriamente los clientes de cada una de las prioridades para no favorecer la priorización actual o la propuesta.

Por lo tanto, los ejecutivos observarán el nombre de sus clientes y la prioridad que tiene cada uno de ellos (de 1 a 7), sin saber si esa prioridad es por la priorización actual o la priorización propuesta.

9 Conclusiones, recomendaciones y propuestas de trabajo futuro

El objetivo principal de este trabajo es desarrollar una metodología para aumentar la tasa de respuesta de las campañas de promoción de productos de consumo pre aprobados y así aumentar las colocaciones de consumo. debido a que, uno de los focos del banco para los próximos años es aumentar el crecimiento de estas colocaciones.

En vista de que en la promoción de estos productos no se realiza asignación exclusiva de canales y se utilizan todos los canales para contactar a los clientes, se decide intervenir en el proceso de priorización de los ejecutivos, ya que, se observa una priorización basada en reglas básicas a partir de los conocimientos de los expertos en el negocio, sin embargo, sin una base estadística de respaldo.

Gracias al análisis descriptivo, se pudo observar que al entregar una priorización a los ejecutivos se aumenta la gestión de los clientes con mayor prioridad, por ejemplo, en la campaña del tercer trimestre del año 2017 la prioridad 1 tuvo un 95% de gestión y la última prioridad tuvo un porcentaje de gestión del 55%, se puede observar que hubo un aumento de 40 puntos porcentuales, por otro lado, si se realiza el mismo ejercicio, también se observa una tasa de respuesta mayor en los clientes con mayor prioridad, ya que, en las prioridades altas (1, 2, 3 y 4), la tasa en promedio es de 8% y en las prioridades bajas (5, 6 y 7) de 3% (ver Tabla 2), se puede observar un aumento de 5 puntos porcentuales. Por lo tanto, el trabajo se basa en la hipótesis de que, al mejorar la priorización, entonces la tasa de respuestas de las campañas eventualmente también mejorará.

El trabajo se abordó desde la metodología CRISP-DM. Para desarrollar satisfactoriamente esta metodología, se recomienda encarecidamente llevar a cabo los primeros pasos con rigurosidad, ya que, es fundamental entender los objetivos del negocio, los procesos involucrados y la información disponible para tener un proyecto exitoso.

Uno de los objetivos específicos corresponde a la construcción de dos modelos de propensión, tanto para la compra de créditos de consumo pre aprobados como para la toma de aumentos de cupo u obtención de tarjeta de crédito pre aprobada. Este objetivo se cumple satisfactoriamente, y al observar los resultados obtenidos, se puede concluir que los modelos aportan información valiosa para identificar a los clientes que toman estos productos y es fundamental dar a conocer a los equipos comerciales el conocimiento adquirido para que puedan incorporarlo dentro de la estrategia comercial, en primer lugar, se observa que los clientes que tienen un fuerte lazo con el banco, es decir, utilizan los servicios y productos que ofrece el banco, son más propensos a utilizar o tomar otra vez un producto con el banco, en particular, los clientes que ya poseen deudas de consumo con el banco son los clientes más propensos a tomar nuevamente un crédito de consumo pre aprobado. Por otro lado, el factor que más peso tiene en la propensión de tarjetas de crédito es la tenencia de tarjeta de crédito vigente,

es decir, los clientes son más propensos a tomar un aumento de cupo que a realizar la apertura de una tarjeta de crédito.

El modelo de propensión a créditos de consumo obtiene resultados satisfactorios al compararlos con la priorización actual. En la muestra de validación se captura el 71% de la venta en el 28% de los clientes más propensos, esto implica una mejora de 23,7 puntos porcentuales, también se alcanza una precisión de 10% en el 28% de los clientes más propensos, lo que significa una mejora de 4,4 puntos porcentuales.

En cuanto al modelo de propensión a los aumentos o apertura de tarjetas de crédito, también se observan resultados satisfactorios, se logra capturar un 50% de la venta en el 20% de los clientes más propensos, esto equivale a un aumento de 14,4 puntos porcentuales con respecto a la priorización actual, así mismo, se obtiene una precisión del 10% en el 20% de los clientes más propensos, lo que implica una mejora de 3,2 puntos porcentuales.

Con respecto a la evaluación de los modelos de propensión, se recomienda definir desde un principio las métricas de evaluación que se desean maximizar, estas medidas de desempeño dependen del problema que se busca solucionar, por ejemplo, los resultados en cuanto a la precisión de los modelos de propensión no son altos, debido al gran desbalance de los datos, pero sí lo son en cuanto al área bajo la curva de ROC, por lo tanto, permiten realizar un buen ranking de clientes para la optimización y cumplir con el objetivo del proyecto, pero no para conocer la probabilidad de que un cliente tome los productos con exactitud.

Por otro lado, el segundo objetivo específico consta de segmentar a los clientes del banco, de acuerdo con el uso de sitio web y correo electrónico. También se cumple con este objetivo y se identifican 4 segmentos, tanto para el uso del sitio web como para el uso del correo electrónico.

Sobre el conjunto de clientes que ha utilizado el sitio web durante los últimos 3 meses antes de una campaña, es decir, dejando fuera a los clientes inactivos, se identifica:

- El 25% como un grupo de clientes de “uso alto”. Estos utilizan este medio frecuentemente y en promedio ingresan al sitio web con su usuario 15 veces mensualmente y no lo han dejado de utilizar.

- El 55% como un grupo de clientes de “uso medio”. Estos utilizan este medio con menor frecuencia, en promedio ingresan al sitio web con su usuario 6 veces mensualmente y tampoco lo han dejado de utilizar.

- El 15% como un grupo de clientes de “uso bajo”. Estos utilizan este medio con mínima frecuencia y en promedio ingresan al sitio web con su usuario 2 veces mensualmente y tampoco lo han dejado de utilizar.

- El 5% como un grupo de “uso muy bajo”. Estos también han utilizado el sitio web en promedio 2 veces mensualmente, sin embargo, la última conexión fue hace más de 2 meses, por lo tanto, estos clientes dejaron de utilizar el sitio web o lo dejaran.

De igual manera, se observa el comportamiento de los clientes en los últimos 6 correos enviados en los últimos 6 meses y también dejando fuera a los clientes a los que no se les envío ningún correo, se identifica:

- El 20% como un grupo de clientes de “uso constante alto”. Estos abren el 78% de los correos en promedio, de los últimos 6 correos enviados en los últimos 6 meses.

- El 16% como un grupo de clientes de “uso reciente alto”. Estos abren el 82% de los correos en promedio, sin embargo, de los últimos 3 correos enviados en los últimos 6 meses.

- El 37% como un grupo de clientes de “uso constante bajo”. Estos abren el 11% de los correos en promedio, de los últimos 6 correos enviados en los últimos 6 meses.

- El 27% como un grupo de clientes de “uso reciente bajo”. Estos abren el 6% de los correos en promedio, sin embargo, de los últimos 3 correos enviados en los últimos 6 meses.

Analizando los segmentos encontrados en conjunto, se observa que el uso del sitio web es más habitual que el uso del correo electrónico, ya que, el 80% de los clientes hace un uso medio o alto del sitio web, a diferencia del uso del correo, donde el 64% de los clientes activos hace un uso bajo del correo. Al realizar la intersección del análisis de uso de canales, existe una gran cantidad de clientes que no usa el sitio web y que tampoco abre los correos electrónicos (14% en promedio), en conclusión, es importante considerar que si estos clientes no son gestionados a través del ejecutivo no se enteraran de los productos pre aprobados que tienen disponibles, es más, en la campaña de agosto de 2017 se usan los SMS para reforzar este tipo de clientes, por lo tanto, el trabajo agrega valor al incorporar este comportamiento en la optimización.

El tercer objetivo específico tiene relación con la propuesta de una priorización, lo cual se cumple al resolver un problema de optimización lineal entero que considera:

- El producto ofertado. En el problema se toma en cuenta por separado el monto ofertado en cada uno de los productos, sin embargo, por la naturaleza de los créditos de consumo, el monto ofertado tiende a ser mayor en este producto, esto se soluciona al incorporar un transformador monetario.

- La probabilidad de compra. El problema toma en cuenta el monto esperado máximo que el cliente puede tomar en cada uno de los productos, al realizar la multiplicación de la probabilidad por el monto ofertado en cada uno de los productos.

- El perfil de uso de canales del cliente. A través de la segmentación de uso del sitio web y correo electrónico se identifica a los clientes que no se enterarán de la oferta, debido a que, constantemente no abren los correos y dejaron de utilizar el sitio web. Por lo tanto, se toma en consideración que si no se gestiona a través de los ejecutivos a estos clientes, entonces se espera que no tomen la oferta, es decir, utilidad esperada nula.

Como propuesta de trabajo futuro:

- Incluir en la optimización la selección de canal, debido a que, se puede aumentar la tasa de respuesta, mejorar la percepción del cliente acerca del banco y reducir los costos de cada campaña

- Incluir una estimación del monto que el cliente va a tomar, debido a que, la estimación de la utilidad esperada en este trabajo corresponde al monto máximo que el cliente puede tomar

- Incluir dentro de los modelos de propensión un indicador de principalidad, debido a que, uno de los factores que se repite en los resultados de estos, es que los clientes cercanos al banco son los más propensos.

- Realizar el experimento para validar el aumento de la tasa de respuesta de las campañas y el aumento de las colocaciones de consumo

10 Bibliografía

1. ABIF. (2016). *La banca centrada en las personas. Memoria 2016.* Obtenido de www.abif.cl

2. Amini, M., Rezaeenour, J., & Hadavandi, E. (2015). A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing. *International Journal of Computational Intelligence and Applications, 14*(4), 23.

3. Ayetiran, E. F., & Adeyemo, A. B. (2012). A Data Mining-Based Response Model for Target Selection in Direct Marketing. *International Journal of Information Technology and Computer Science, 4*(1), 9-18.

4. Batista, G., Bazzan, A., & Monard, M. C. (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study. *WOB*, 10-18.

5. Batista, G., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29.

6. Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization.* Belmont, MA: Athena Scientific.

7. Bose, I., & Chen, X. (16 de Mayo de 2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research, 195*(1), 1-16.

8. Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications, 42*(22), 8403-8412.

9. Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

10. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Ai Magazine, 17*(3), 37.

11. Guyon, I., Weston, J., Barnhill, S., & Vladimir, V. (2002). Gene Selection for Cancer Classification. *Machine learning, 46*(1), 389-422.

12. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*(2), 107-145.

13. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.* Elsevier.

14. Knott, A., Hayes, A., & Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing, 16*(3), 59-75.

15. Krejcie, R. V., & Morgan, W. D. (1970). Determining Sample Size for Research Activities. *Educational and Psychological Measurement , 30*(3), 607-610.

16. Lin, A. Z. (2013). Variable Reduction in SAS by Using Weight of Evidence and Information Value. *SAS Global Forum*, 95-213.

17. Linoff, G., & Berry, M. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.

18. Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing, 15*(5), 494-504.

19. SBIF. (2017). *Panorama Bancario Primer Trimestre 2017.* Obtenido de www.sbif.cl

20. Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5*(4), 13-22.

21. SINACOFI. (2017). *SINACOFI: Nuestra empresa*. [En línea] <https://www.sinacofi.cl/nuestra_empresa.asp> [12 de Junio de 2017]

22. Triola, M. F. (2017). *Elementary statistics.* Boston: Pearson.

23. WEF. (2016). *The Global Competitiveness Report 2016-2017.* Obtenido de www.weforum.org

24. Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management, 4*(19), 4199-4206.

11 Anexos

Anexo A:

Tabla 27: Resultados de ventas de tarjetas de crédito por prioridad en campaña de julio, agosto y septiembre de 2017

Prioridad	Porcentaje del total de ventas	Tasa de respuesta	Porcentaje de clientes
1	**0,3%**	**2,7%**	**0,4%**
2	**2,5%**	**5,3%**	**1,8%**
3	**15,2%**	**9,3%**	**6,2%**
4	**17,7%**	**5,8%**	**11,6%**
5	49,4%	3,8%	48,8%
6	5,0%	2,7%	6,9%
7	9,9%	1,6%	24,3%
Total	**100%**	**3,8%**	**100%**

Fuente: Elaboración propia.

Anexo B:

Tabla 28: Información seleccionada y formato de la fuente.

Nombre de la base de datos	Formato	Información seleccionada
CIF	Estático	ID Tipo de persona: natural o jurídica Tipo de cliente: empresa, corporación o persona Segmento de la banca personas Clasificación comercial
Demográficos	Estático	ID Sexo Profesión Estado civil
Vehículos	Estático	ID Avalúo Año de fabricación
Bienes raíces	Estático	ID

		Avalúo
Activos	Mensual	ID Fecha Monto comprobado de bienes raíces Monto comprobado de inversiones Monto comprobado de otros activos Monto comprobado de participaciones en sociedades Monto comprobado de vehículos
Pasivos	Mensual	ID Fecha Monto comprobado en tarjeta de crédito Monto comprobado en línea de sobregiro Monto comprobado en obligaciones a corto plazo Monto comprobado en obligaciones a largo plazo Monto comprobado en leasing Monto comprobado en otros pasivos Monto comprobado de patrimonio
Tenencia de productos	Mensual	ID Tipo de producto
Apertura de productos	Transaccional	ID Fecha Tipo de producto
Deuda en el Banco	Mensual	ID Fecha Deuda consumo Deuda comercial Deuda hipotecaria Línea de crédito
Deuda informada por la SBIF	Mensual	ID Fecha Deuda consumo Deuda comercial

		Deuda hipotecaria Línea de crédito
Saldos vista	Mensual	ID Saldo vista promedio
Sitio web	Clickstream (registro del recorrido que sigue el usuario por las páginas del sitio web)	ID Fecha Tiempo de visita
Aplicación	Diario	ID Fecha Cuenta de visitas diarias
Correos electrónicos	Transaccional	ID Cantidad de aperturas Tipo de campaña Fecha de primera apertura Fecha de última apertura
Call Center	Transaccional	ID Fecha Tipo de llamada
Simulaciones de créditos pre aprobados	Transaccional	ID Fecha
Transferencias electrónicas	Transaccional	ID Fecha Monto Banco de origen Banco de destino Tipo: abono o cargo Misma persona (si el destino de la transferencia es a la misma persona)
Pago automático a cuenta corriente	Transaccional	ID Fecha Monto
Pago automático a tarjeta de crédito	Transaccional	ID Fecha Monto
Tarjeta de débito	Transaccional	ID Identificador de tarjeta Tipo: giro o pago Fecha

		Monto Rubro
Tarjeta de crédito	Transaccional	ID Identificador de tarjeta Fecha Monto Cuotas Rubro
Cupos y utilización de línea de sobregiro	Mensual	ID Fecha Fecha cierre de producto Identificador de línea de sobregiro Cupo línea de sobregiro Monto utilizado
Cupo y deuda de tarjeta de crédito	Diario	ID Fecha Cupo nacional Cupo internacional Deuda nacional Deuda internacional
Campaña Pre aprobados	Mensual	ID Fecha Indicador de venta

Fuente: Elaboración propia.

Anexo C:

Tabla 29: Análisis descriptivo antes y después de realizar filtros de valores atípicos

	PAC ($)		Pagos con débito ($)		Giros con débito ($)	
	Sin filtro	Con filtro	Sin filtro	Con filtro	Sin filtro	Con filtro
Porcentaje	100%	98,2%	100%	99,6%	100%	99,5%
Datos	1.094.824	1.074.952	9.284.513	9.247.222	2.544.269	2.531.398
Promedio	826.885	418.761	22.451	21.632	59.013	57.491
D. Estándar	37.421.257	666.407	34.624	29.806	63.541	59.858
Mínimo	10	10.000	1	500	524	2.000
25% (Q1)	50.000	50.000	5.000	5.000	15.000	15.000
50% (Q2)	199.900	187.405	11.750	11.716	30.000	30.000
75% (Q3)	508.900	498.000	26.180	26.003	80.000	80.000
Máximo	2,39E+10	5.000.000	500.000	300.000	400.000	200.000

	TEF Abonos ($)		TEF Cargo ($)		PAT ($)	
	Sin filtro	Con filtro	Sin filtro	Con filtro	Sin filtro	Con filtro
Porcentaje	100%	96,7%	100%	97,7%	100%	98,3%
Datos	2.009.910	1.942.080	3.218.531	3.144.600	979.246	962.563
Promedio	465.855	333.157	326.424	237.165	54.619	38.081
D. Estándar	911.583	546.228	753.534	442.135	310.966	63.452
Mínimo	1	500	1	1.000	1	500
25% (Q1)	25.000	24.000	25.000	25.000	5.518	5.403
50% (Q2)	100.000	99.999	75.000	70.091	20.000	19.427
75% (Q3)	447.167	388.287	250.000	220.000	45.000	43.658
Máximo	7.000.000	3.000.000	7.000.000	3.000.000	46.695.365	500.000

	Tarjeta de crédito ($)	
	Sin filtro	Con filtro
Porcentaje	100%	97,7%
Datos	9.993.856	9.762.698
Promedio	64.635	35.138
D. Estándar	658.043	52.506
Mínimo	1	500
25% (Q1)	6.190	6.140
50% (Q2)	17.600	17.000
75% (Q3)	41.789	39.990
Máximo	400E+6	400.000

Fuente: Elaboración propia.

Anexo D:

Tabla 30: Variables seleccionadas para el modelo de propensión de créditos de consumo

Descripción
Porcentaje de giro en tarjeta débito de los últ. 6 meses
Ratio entre el monto total TEF abonado de los últ. 3 meses sobre últ. 12 meses
Posee crédito de consumo que vence en menos de 6 meses
Porcentaje de compras con cuotas en tarjeta de crédito de los últ. 12 meses
Ratio entre la deuda de consumo SBIF promedio de los últ. 3 meses sobre últ. 6 meses
Días desde la última simulación de crédito de consumo
Porcentaje de cargos TEF a misma persona en los últ. 12 meses
Realiza apertura de consumo en los últ. 12 meses
Saldo vista mensual promedio del últ. Año
Deuda de consumo promedio en Banco de los últ. 12 meses
Monto de inversiones comprobado al últ. mes

Fuente: Elaboración propia.

Anexo E:

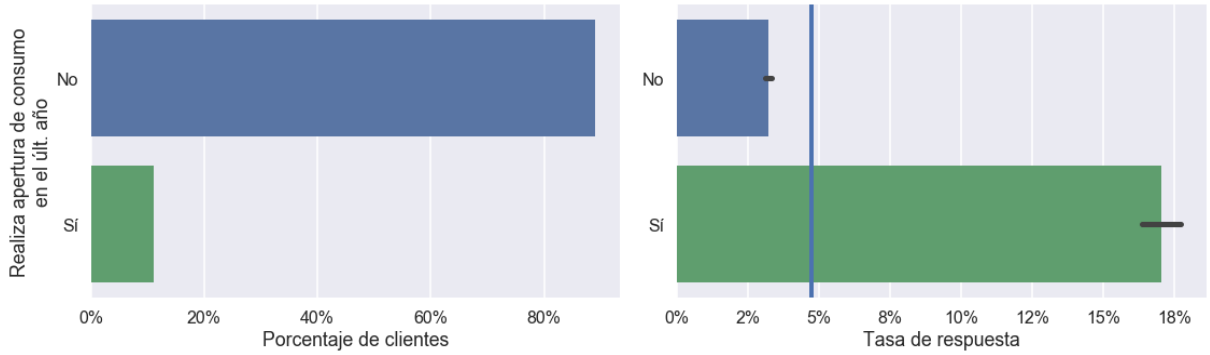
Tabla 31: Variables seleccionadas para el modelo de propensión de tarjetas de crédito

Descripción
Ratio entre la línea de crédito disponible promedio de los últ. 3 meses sobre los últ. 12 meses
Género
Saldo vista promedio de los últ. 12 meses
Deuda de consumo máxima en Banco de los últ. 12 meses
Porcentaje de compras realizadas con cuotas en los últ. 6 meses
Tenencia de tarjeta de crédito vigente
Edad

Fuente: Elaboración propia.

Anexo F:

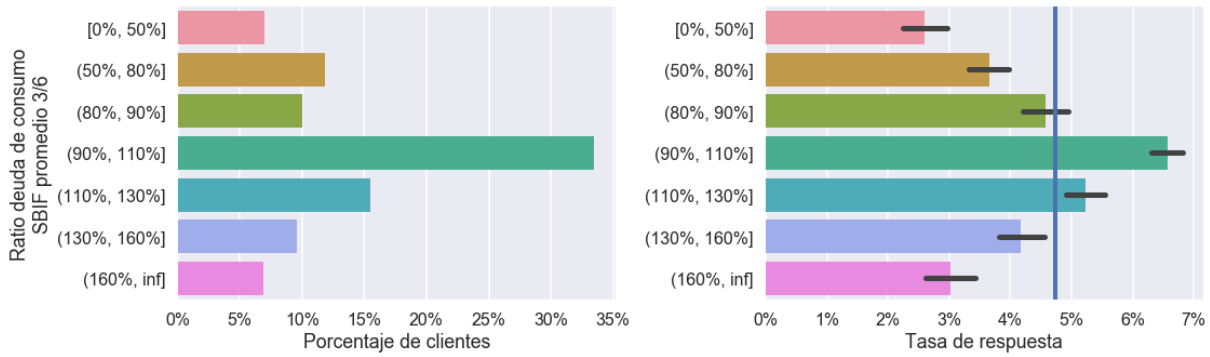
Ilustración 42: Porcentaje de clientes y tasa de respuesta de crédito de consumo por indicador de si el cliente ha realizado una apertura de consumo en el último año



Fuente: Elaboración propia.

Anexo G:

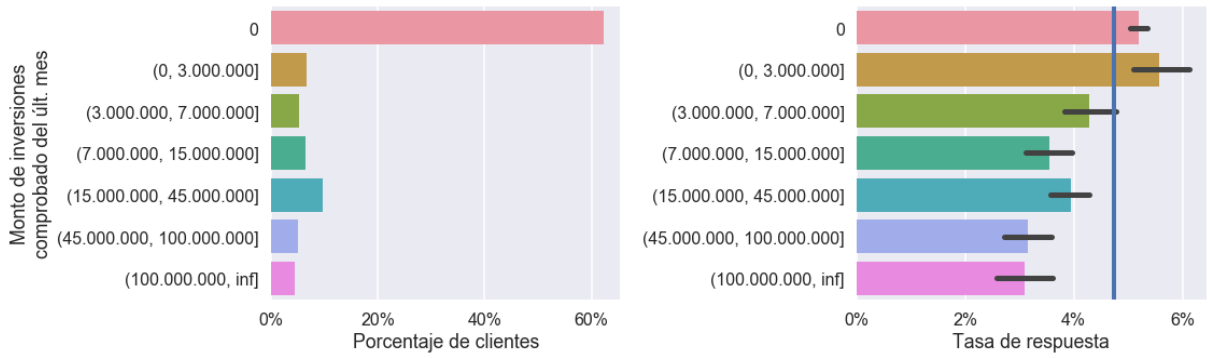
Ilustración 43: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del ratio del promedio de la deuda de consumo en el banco de los últimos tres meses sobre los últimos 6 meses



Fuente: Elaboración propia.

Anexo H:

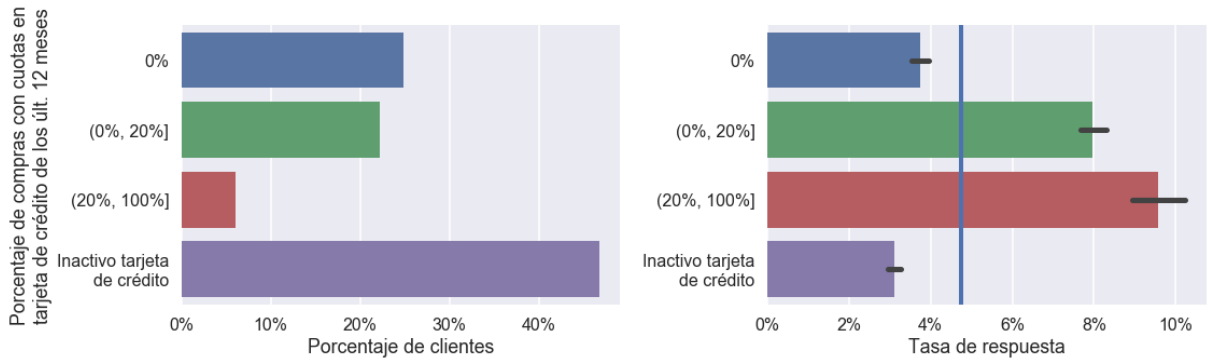
Ilustración 44: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del monto de inversiones comprobado del último mes



Fuente: Elaboración propia.

Anexo I:

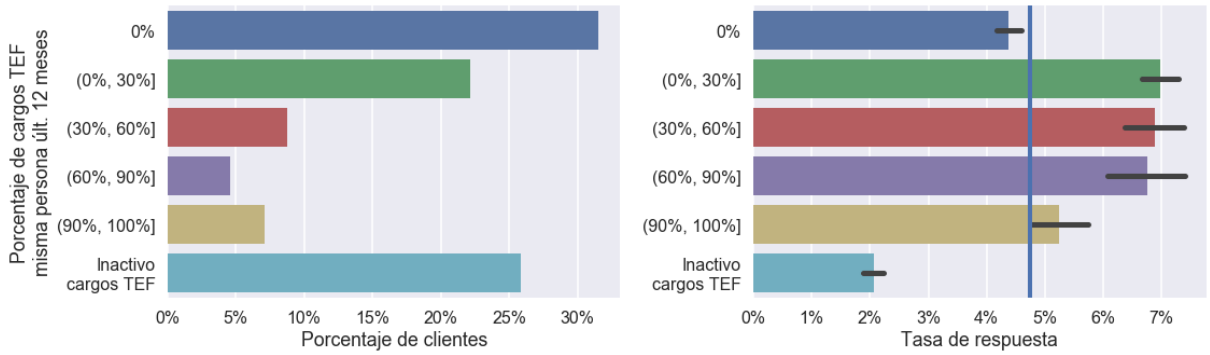
Ilustración 45: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del porcentaje de compras con cuotas en tarjeta de crédito de los últimos 12 meses



Fuente: Elaboración propia.

Anexo J:

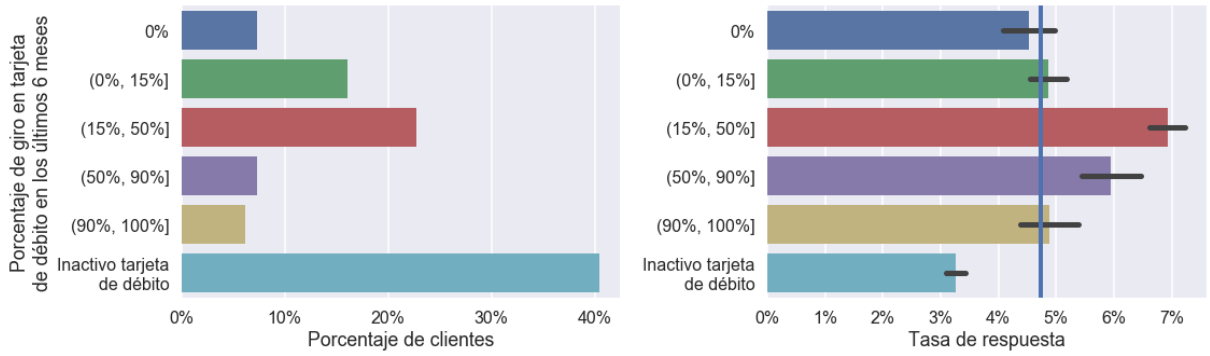
Ilustración 46: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del porcentaje de cargos en transferencias electrónicas a la misma persona en los últimos 12 meses



Fuente: Elaboración propia.

Anexo K:

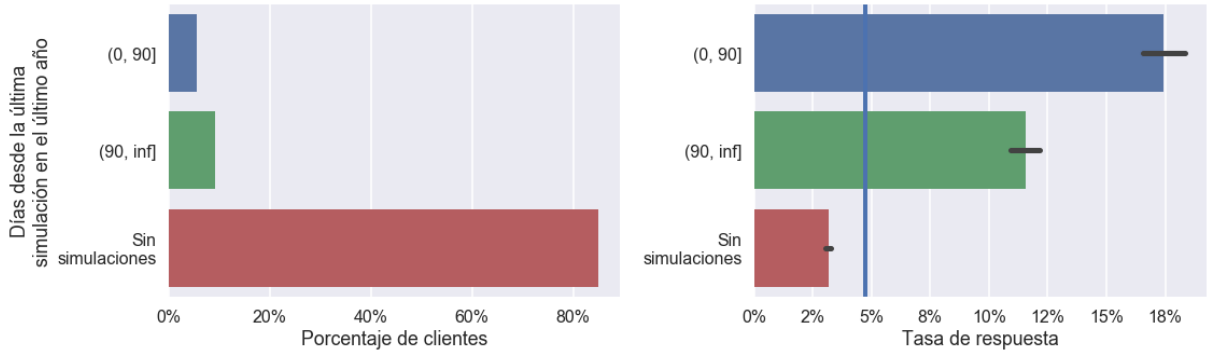
Ilustración 47: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del porcentaje de giros en tarjeta de débito de los últimos 6 meses



Fuente: Elaboración propia.

Anexo L:

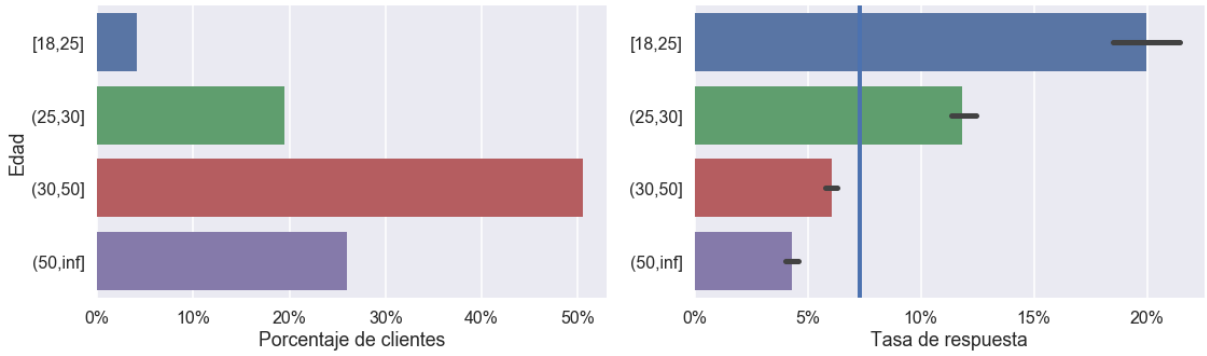
Ilustración 48: Porcentaje de clientes y tasa de respuesta de crédito de consumo por tramo del porcentaje del ratio entre el monto total abonado en transferencias electrónicas de los últimos 3 meses sobre los últimos 12 meses



Fuente: Elaboración propia.

Anexo M:

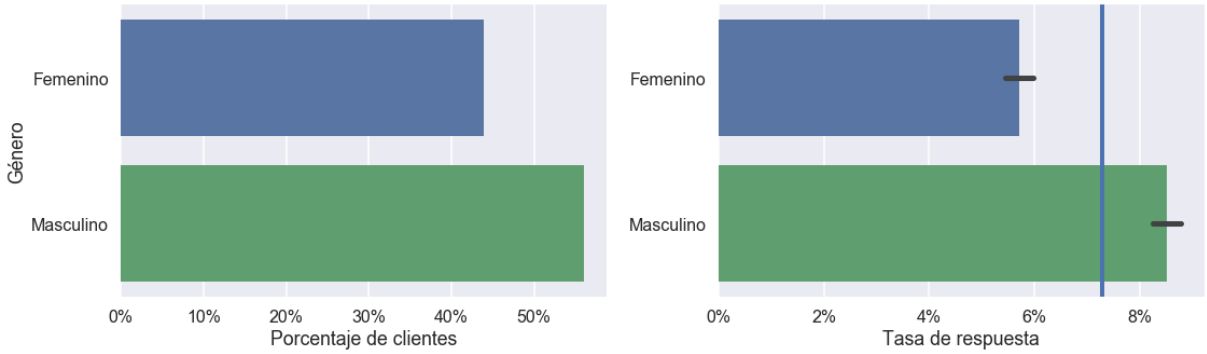
Ilustración 49: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por tramo de edad



Fuente: Elaboración propia.

Anexo N:

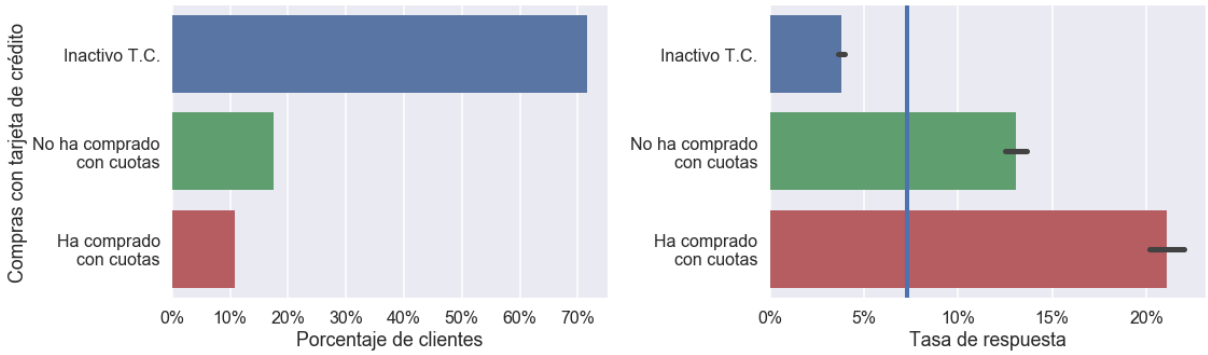
Ilustración 50: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por género



Fuente: Elaboración propia.

Anexo O:

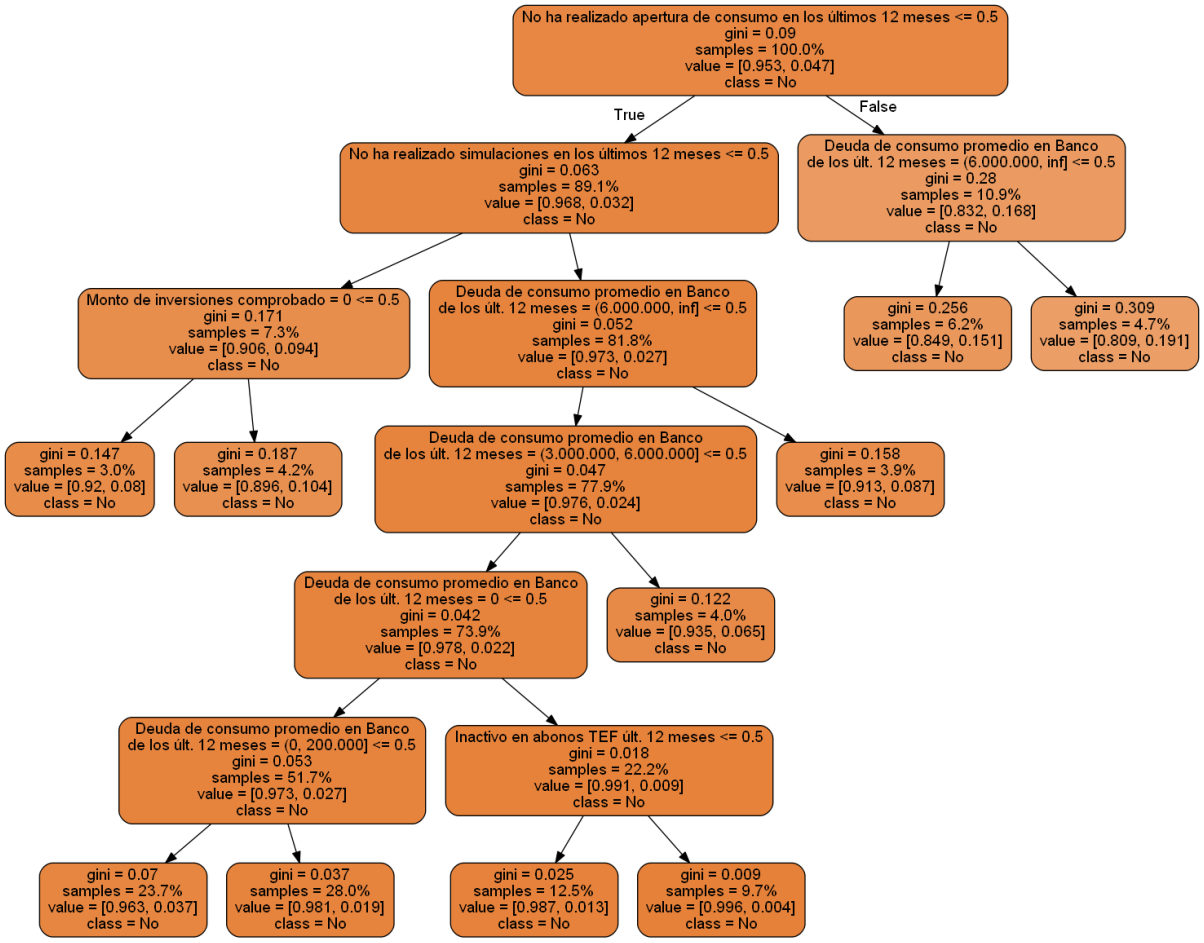
Ilustración 51: Porcentaje de clientes y tasa de respuesta de aumento de tarjeta de crédito por compras con tarjeta de crédito en los últimos 6 meses



Fuente: Elaboración propia.

Anexo P:

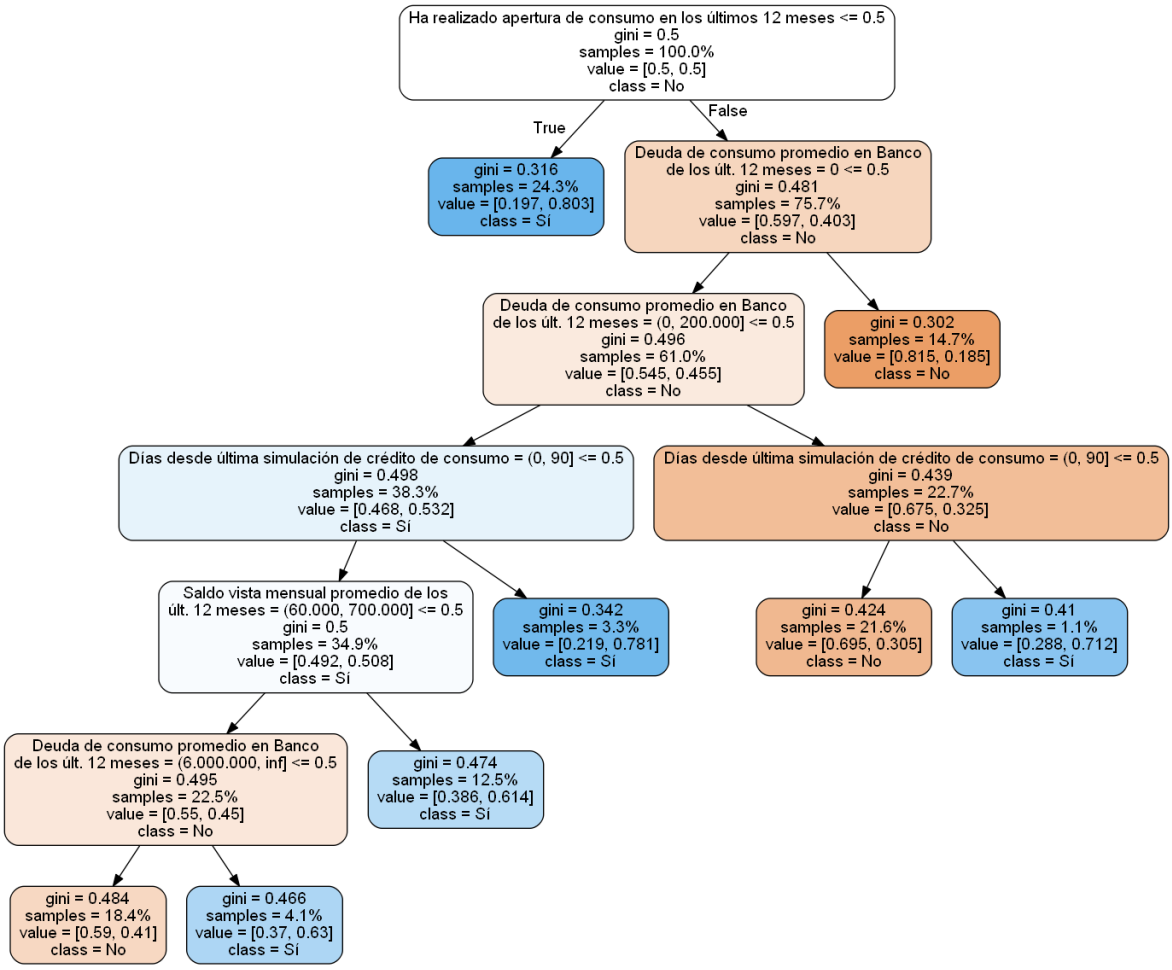
Ilustración 52: Árbol de decisión de modelo de propensión de compra de créditos de consumo sin método de balanceo



Fuente: Elaboración propia.

Anexo Q:

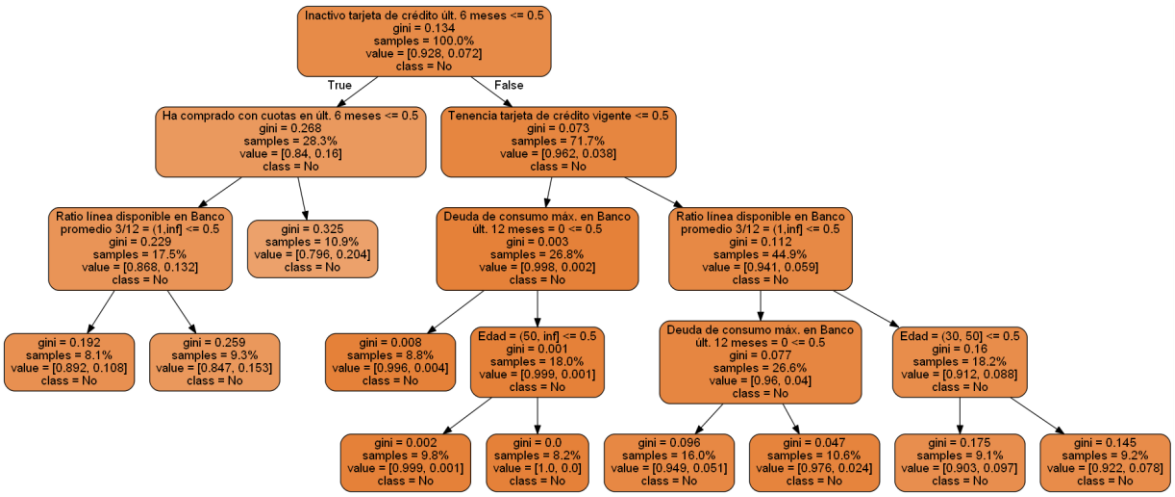
Ilustración 53: Árbol de decisión de modelo de propensión de compra de créditos de consumo con método SMOTE + Tomek.



Fuente: Elaboración propia.

Anexo R:

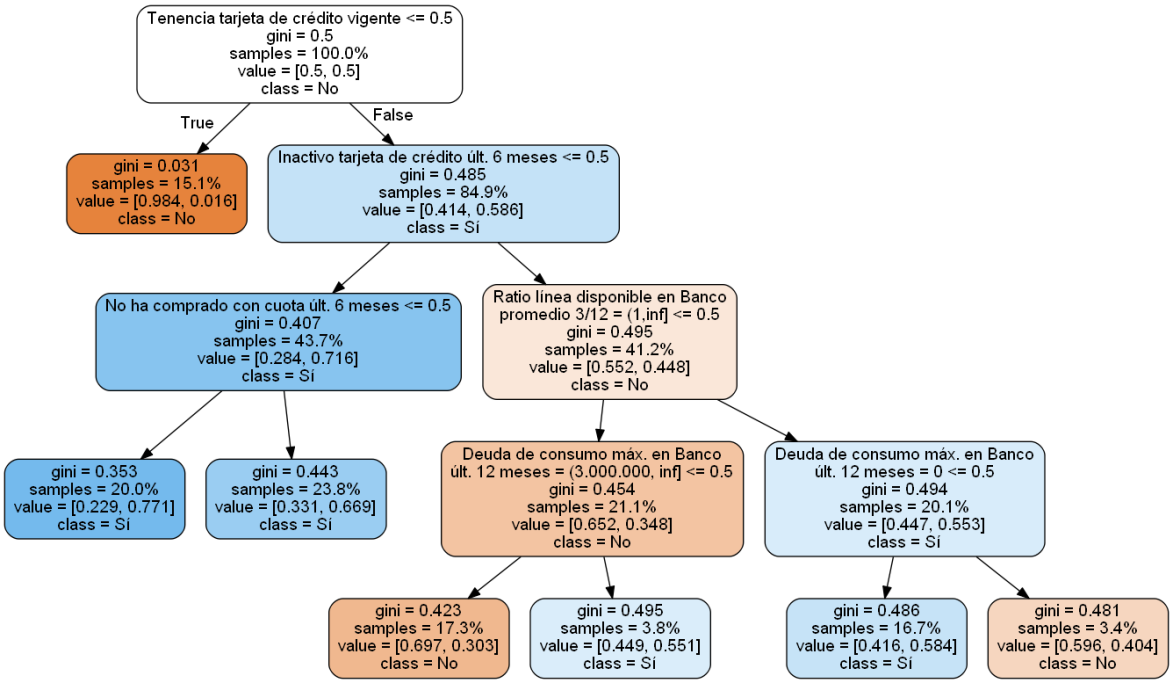
Ilustración 54: Árbol de decisión de modelo de propensión de compra de tarjetas de crédito sin balanceo de datos



Fuente: Elaboración propia.

Anexo S:

Ilustración 55. Árbol de decisión de modelo de propensión de compra de tarjetas de crédito con método SMOTE + Tomek



Fuente: Elaboración propia.