UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE FUGA CON MINERÍA DE TEXTO EN DIFERENTES CANALES PARA EMPRESA DE RETAIL FINANCIERO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JUAN ENRIQUE ALDUNATE CASTILLO

PROFESOR GUÍA:

SEBASTÍAN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:

CAROLINA SEGOVIA RIQUELME

ERICK MÉNDEZ GUZMÁN

SANTIAGO DE CHILE

2018

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Juan Enrique Aldunate Castillo
FECHA: 10/04/2018
PROFESOR GUÍA: Sebastián Ríos Pérez

MODELO DE FUGA CON MINERÍA DE TEXTO EN DIFERENTES CANALES PARA EMPRESA DE RETAIL FINANCIERO

El presente informe contiene el análisis realizado para una empresa de retail financiero, y considera variables obtenidas por medio de minería de texto para incluirlas dentro de dicho modelo.

Dentro de la empresa estudiada, se tienen dos tipos de clientes. Uno de ellos suele realizar comentarios con la empresa, tanto positivos como negativos. Estos además suele gastar más en un periodo de tiempo (en promedio) que un cliente que no lo realiza. Por ejemplo, dentro de un determinado mes el cliente que realiza comentarios gasta aproximadamente entre dos y tres veces más que su contraparte que no realiza comentarios. Al mismo tiempo, los clientes que si realizan comentarios son tres veces menos propensos a la fuga que los que no los realizan. Ante esto surge la oportunidad de negocios de determinar si por medio de herramientas de minería de texto se puede extraer información útil de los comentarios para predecir la fuga de los clientes, y minimizarla con estas variables.

El objetivo de esta memoria es construir un modelo predictivo de fuga utilizando textos para poder mejorar la retención de clientes.

Para lograr cumplir con este objetivo, se realizaron 4 modelos de fuga. El primer modelo de fuga consistía en un modelo de fuga tradicional (accuracy:77,5% ,presicion:82% y AUC:85,1%), un modelo de fuga con variables de minería de texto solo para los clientes que tienen comentario (accuracy: 77,9% ,presicion:80,9% y AUC: 85,4%), un modelo de fuga con variable de minería de textos para toda la cartera sin imputar valores desconocidos (accuracy: 76,1% ,presicion:97,7% y AUC:83,7%) y un modelo de fuga con variable de minería de texto imputados para aquellos que no tenían comentario (accuracy: 75,7% ,presicion:82% y AUC:84,2%).

Con respecto a los resultados del modelo de fuga, se concluye que dentro de todos los modelos estudiados, el mejor modelo es uno que considera variables de minería de texto sobre una base de clientes en que todos emiten comentarios. Dado que este no se puede generalizar para todos los clientes de la empresa, decide utilizar el modelo de fuga tradicional sin variables de minería de texto. Otra conclusión importante con respecto a la memoria es que la imputación de las variables de minería de textos no causa un aumento en el rendimiento con respecto a un modelo que no lo hace. Esto se debe a que al comparar los modelos con y sin imputación de datos, el modelo que incluye la imputación de los datos de la variable de minería de texto tiene menor accuracy (0,4% menos de accuracy), presenta un AUC levemente mayor (0,5% más de AUC). A pesar de esto, las variables de minería de texto apoyan dentro de la gestión de los clientes que si realizan comentarios.

AGRADECIMIENTOS

Agradezco a mi profesor guía por todas las horas de apoyo al desarrollo de mi memoria. Siento que fue un verdadero aporte a la creación de mi memoria y de que la calidad del trabajo mejoró considerablemente una vez que lo tuve por guía.

Agradezco también a mi familia y amigos por ayudarme en las diferentes etapas de la memoria. Sin su apoyo creo que no habría sido posible la confección del trabajo que logre, por lo que les estoy muy agradecidos por todo.

Por ultimo agradezco a mis compañeros de trabajo. Sin su buena disposición a responder consultas y a apoyarme a lo largo del proceso creo que no se habría podido lograr lo que se logró de buena manera.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. ANTECEDENTES GENERALES

1.1 Industria de servicios financieros

La compañía estudiada en esta memoria pertenece al rubro de servicios financieros. Estas se definen como empresas encargadas del préstamo y operación de créditos, usualmente por medio de tarjetas de crédito y otras herramientas. Sus principales actores dentro del país se presentan en la siguiente tabla.

	Empresa	Empresas de retail/tiendas asociadas
Entidades bancarias	Banco Santander	
	Banco de Chile	
	Banco del estado de chile	
	Banco de Crédito e Inversiones	
	Banco Bilbao Vizcaya Argentaria, Chile (BBVA)	
	Banco Itaú corpbanca	
	Banco Scotiabank Chile	
	Banco Security	
	Banco BICE	
	Banco Internacional	
Entidades no bancarias	LP S.A.	
	Coopeuch	
	Consorcio tarjetas de crédito	
	Promotora CMR	Falabella
	CAR S.A.	Ripley
	CAT Administradora de tarjetas S.A.	Cencosud
	Presto S.A.	Líder
	Crédito organización y finanzas S.A.	Abcdin, Dijon
	Inversiones y tarjetas S.A.	Hites

Ilustración 1: Principales actores industria de servicios financieros
Fuente: Elaboración propia con datos de la Superintendencia de Bancos e Instituciones financieras (SBIF) y Superintendencia de Valores y Seguros (SVS)

Con respecto a la evolución de la industria dentro de los últimos años, esta evolución puede ser apreciada en el siguiente gráfico:

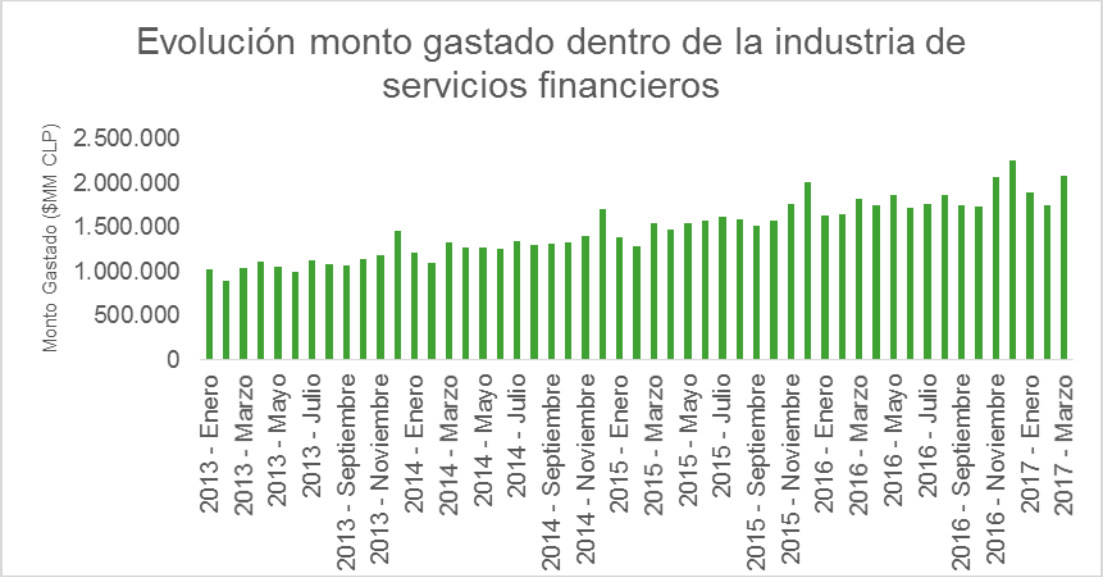


Grafico 1: Evolución del gasto dentro de la industria de servicios financieros
Fuente: Elaboración propia con datos de la Superintendencia de Bancos e Instituciones financieras (SBIF) y Superintendencia de Valores y Seguros (SVS)

Tal como se observa en el gráfico, esta es una industria que tiene una tendencia de crecimiento sostenida en los últimos meses de estudio, con lo que el aumento de participación de mercado se vuelve más importante, para poder así consolidar dicha posición cuando esta industria llegue a su estado estacionario.

Las instituciones incluidas en esta categoría se dividen en dos grupos: entidades bancarias y no bancarias. Una de las diferencias importantes a destacar entre ambos tipos es el monto de interés que estas empresas están permitidas a cobrar a sus clientes. Los principales organismos reguladores dentro de la industria son la superintendencia de bancos e instituciones financieras (SBIF) y la superintendencia de valores y seguros (SVS).

Las entidades no bancarias suelen estar asociadas a alguna casa comercial, dado que la generación de una tarjeta de crédito constituye una alternativa para fomentar las ventas en estas instituciones. Este es el caso para esta empresa, dado que pertenece a uno de los principales holdings del país, con filiales de retail, supermercados y artículos de hogar.

La empresa a partir de la cual se realizó el presente informe es una entidad no bancaria. Mientras la empresa debió ingresar a la SBIF el año 2006, para poder emitir y operar tarjetas de crédito no bancarias -dado que este organismo es el encargado de velar por el correcto funcionamiento de este tipo de productos- , también es parte de la SVS, puesto que debió emitir valores no accionarios que son parte de algunos de sus productos ofertados por la compañía. Las instituciones encargadas de la fiscalización realizan periódicamente revisiones del estado de la compañía.

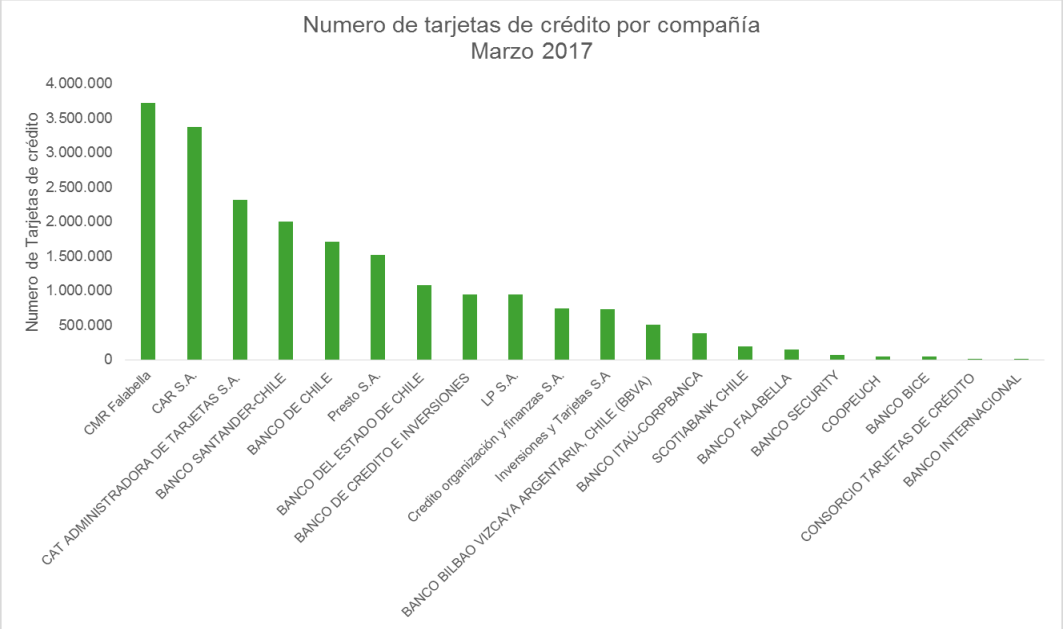


Grafico 2: Numero de tarjetas de crédito por compañía en el año 2016
Fuente: Elaboración propia con datos de la Superintendencia de Bancos e Instituciones financieras (SBIF) y Superintendencia de Valores y Seguros (SVS)

En el grafico se puede apreciar que las empresas de instituciones no bancarias suelen contar con una mayor cantidad de número de tarjetas que las instituciones bancarias. Es importante destacar además que grupos corporativos más poderosos dentro de este ámbito son los de Falabella y Ripley.

Grafico 3: Monto gastado por los clientes con las tarjetas de crédito de diferentes instituciones en el año 2016
Fuente: Elaboración propia con datos de la Superintendencia de Bancos e Instituciones financieras (SBIF) y Superintendencia de Valores y Seguros (SVS)

Al mismo tiempo, se puede estudiar el monto que logran manejar estas tarjetas de crédito por compañía, como se puede ver en el grafico anterior.

A partir de estos datos se puede concluir que pese a que las instituciones bancarias emiten un menor número de tarjetas de crédito que las casas comerciales, son las tarjetas de las instituciones bancarias las que manejan mayores montos gastados.

Utilizando los datos de tarjetas emitidas en el grafico 2 y montos gastados utilizando dichas tarjetas en el gráfico 3, es posible confeccionar un ranking en que el primer lugar lo ocupe la empresa con mayor cantidad de tarjetas y mayores montos gastados.

Siguiendo esta lógica, el primer lugar en las empresas bancarias lo ocupa Banco Santander, y en las empresas no bancarias CMR Falabella. Al comprar entre ambas instituciones, Banco Santander se queda con la primera posición al manejar una cifra mayor de dinero con su tarjeta de crédito.

El principal proveedor de las compañías de esta industria es Transbank. Esta compañía entrega la infraestructura tecnológica a las empresas pertenecientes al rubro, que permite obtener la información de uso que le dan los diferentes clientes de manera rápida y segura. Con esta compañía mantienen una relación bastante lejana. Esto se debe a que, si bien existe envió de comunicación entre las distintas compañías, la mayoría de esta se encuentra automatizada, y comunicación directa no es muy necesaria a menos de existir algún problema de funcionamiento.

1.2 Descripción de la Empresa

La empresa estudiada tiene en particular el objetivo de la emisión de tarjetas de crédito y la realización de todas las actividades complementarias del giro principal. Esta compañía pertenece a uno de los holdings más grandes del país, abordando rubros de retail, mejoramiento del hogar y supermercados. Es importante también destacar que la tarjeta puede ser empleada tanto en las tiendas del holding como en establecimientos externos a él.

Esta compañía pertenece al rubro de servicios financieros, como fue mencionado previamente. Esto se debe principalmente a la operación de una tarjeta de crédito, pero además provee de otras herramientas de financiamiento para sus clientes. Estos son:

- Avances en efectivo: Esta herramienta permite solicitar avances en efectivo en montos monetarios de entre $5.000 a $5.000.000 pesos. Es importante notar que esta cantidad de dinero se obtiene a partir del cupo de la tarjeta, y que además se puede pagar entre 3 y 24 cuotas.

- Súper Avances: Esta herramienta es un cupo adicional al cupo de compra online de tu tarjeta de crédito que va desde $100.000 hasta $6.000.000.Es importante notar además que esto se puede pagar entre 12 y 48 cuotas.
- PAT/PAC: Consiste en una alternativa del pago automático de cuentas, por medio de la tarjeta del holding. Esto se puede hacer para servicios básicos (luz, agua, gas), además de otros servicios como televisión por cable, internet, etc.
- Bip post-pago: Esta alternativa consiste en el pago de transporte público por medio de la tarjeta del holding. Esto tiene el beneficio para los clientes de que con esto el cobro del transporte pasa automáticamente a su cuenta de la tarjeta.

Pese a que esta empresa no tiene un perfil de cliente bien definido, se encuentra orientada al segmento de los mayores de edad, con bajo riesgo financiero. Esta empresa mantiene una relación cercana con los clientes, donde la compañía está siempre interesado en lo que opinan los clientes y en poder mejorar los servicios que brindan como compañía. Este interés de la compañía se puede ver reflejado a través del uso de diferentes plataformas digitales (página web, redes sociales) y de variadas instancias sociales en las que la compañía obtiene constantemente retroalimentación de su servicio.

Un elemento destacable que posee la compañía es de su club de fidelización. Dicho club de fidelización se sustenta en las compras realizadas usando la tarjeta, de manera tal que cada vez que se gastan una cantidad definida de pesos se generan puntos, una herramienta de la compañía que al acumularse pueden ser canjeados por diferentes productos del holding. Existe una gran variedad de canjes posibles, dado que estos puntos se pueden canjear en cualquiera de los comercios que forman parte de este holding.

1.3 Desempeño Organizacional

El siguiente gráfico, que muestra las utilidades anuales de la empresa en los últimos ocho años, facilita la comprensión de la evolución de la misma.

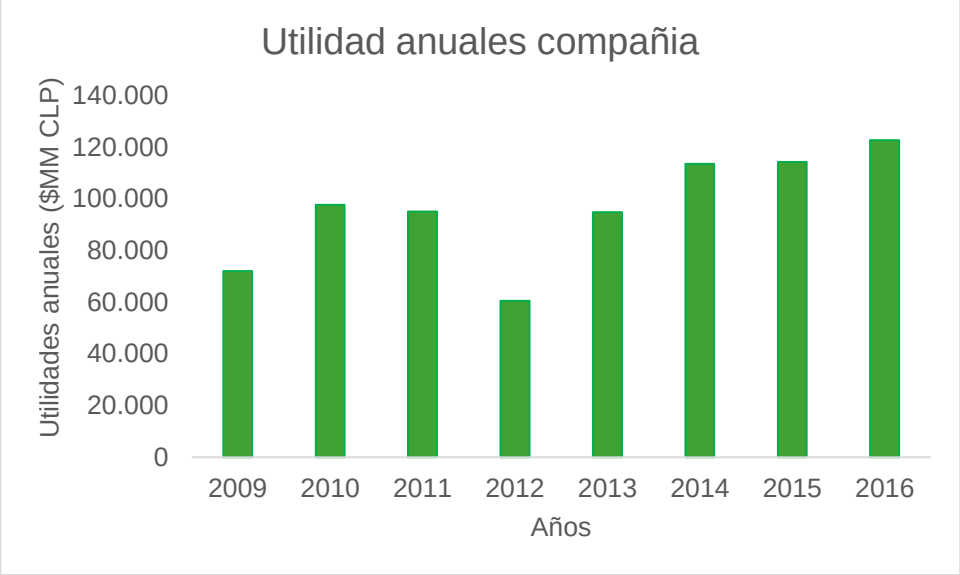


Grafico 4: Utilidades anuales de la empresa
Fuente: Elaboración propia con datos de la empresa

Las utilidades de la empresa estudiada han mantenido una tendencia crecientes desde el año 2012, no obstante, la magnitud de dicho crecimiento es menor a los observados con anterioridad a 2012. De estos datos es posible concluir que la compañía se está aproximando a un estado estacionario, y que en la actualidad su ritmo de crecimiento es menor que en periodos previos.

Asimismo, al analizar el número de clientes activos de la compañía por años, se replica la misma tendencia. Esto apoya la conclusión previa.

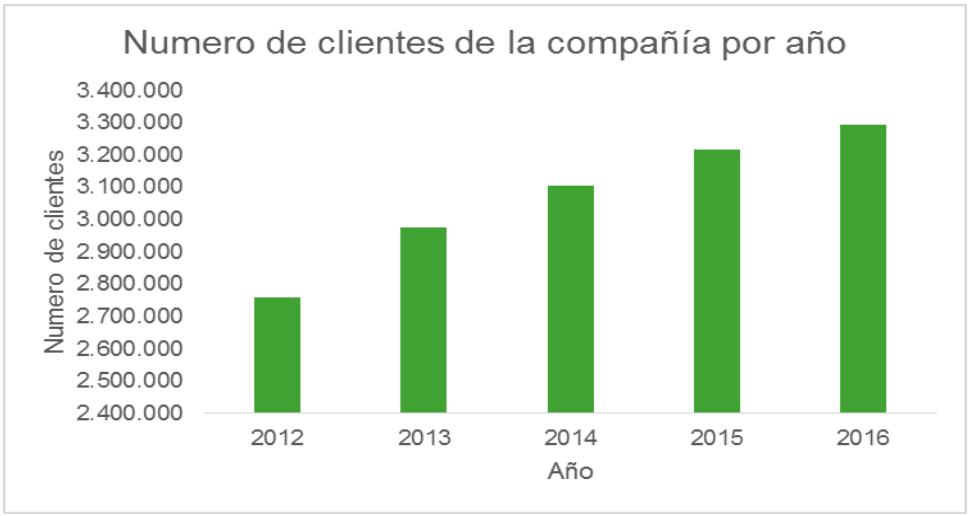


Grafico 5: Número de clientes en la compañía
Fuente: Elaboración propia con datos de la empresa

Otro parámetro para evaluar el desempeño de la empresa en los últimos años es el estudio de la fuga de clientes, evolución que se presenta en el gráfico número 6.

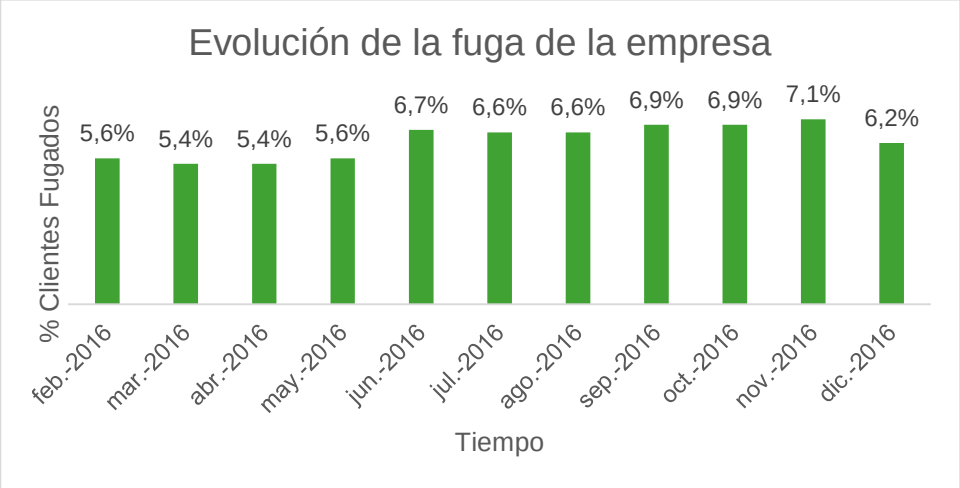


Grafico 6: Evolución del porcentaje de fuga dentro de la compañía
Fuente: Elaboración propia con datos de la empresa

En base a los gráficos presentados anterior se puede ver que la fuga puede es un problema importante en el futuro de la empresa. Esto se debe a que si el número de clientes va aumentando en el tiempo mientras que la tasa de fuga se mantiene constante en el tiempo (en aproximadamente 6,5% de la cartera), esto implica que a medida que avance el tiempo se estarán fugando cada vez más clientes de la compañía. Estos clientes, de ser retenidos podrían tener un efecto considerable en la utilidad de la empresa.

Para poder también entender como ha sido los datos escritos acumulados por parte de la compañía es útil lograr ver la evolución de estas cantidades en el tiempo. Esto se puede apreciar en el siguiente gráfico.

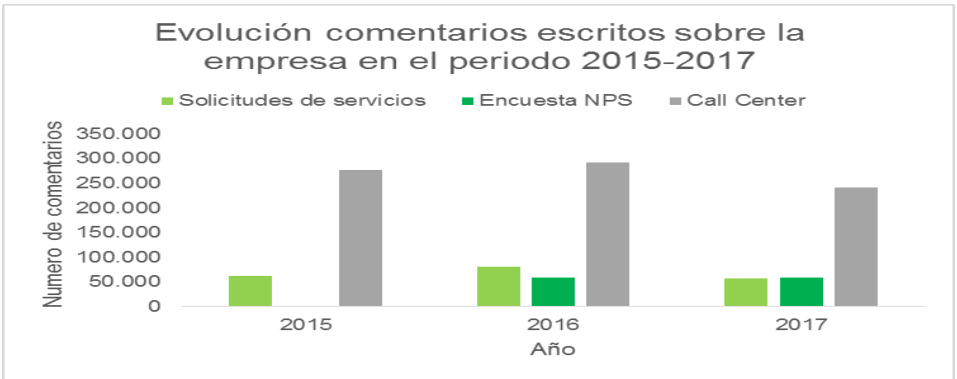


Grafico 7: Numero de interacciones por año
Fuente: Elaboración propia con datos de la empresa

En relación al número de comentarios escritos generados por los clientes, se observa un aumento sostenido durante el periodo de estudio. Pese a que el número de comentarios generados en 2017 es menor que en 2015 y 2016, este registro solamente incluye el primer semestre de 2017 (de enero hasta julio), lo cual permite inferir que el total del año será mayor a los de los años previos.

La importancia de los datos originados en las encuestas a los clientes radica en que representan la opinión de éstos sobre la empresa, por lo cual la detección y posterior corrección de los problemas por ellos referidos puede reducir la fuga de clientes, al mejorar su satisfacción.

2. DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

A partir de los antecedentes generales de la empresa discutidos en la sección anterior, se advierte que la fuga de clientes constituye un importante problema para la compañía y que en los últimos años la retroalimentación recibida desde los clientes ha ido en aumento.

El objetivo de la presente sección es determinar si existe relación entre estos dos aspectos, y analizar si la explotación de estos datos permitiría tomar medidas pertinentes a reducir a fuga de clientes.

2.1 Información de los Datos

En esta memoria se utilizaron dos bases de datos. Estas se presentan en la siguiente tabla.

Base de datos	Descripción
Base de información	Contiene la información de los clientes entre marzo y octubre del año 2016, incluyendo tanto parámetros sociodemográficos como del comportamiento transaccional sobre productos asociados a la empresa
Base de interacciones	Incluye las diferentes interacciones que los clientes han realizado con la compañía, a partir del año 2003

Ilustración 2: Bases de datos utilizados en la memoria
Fuente: Elaboración propia con datos de la empresa

A continuación se presentan los campos de la base de datos de información con sus respectivos descriptores.

Tipo de variable	Descripción	Variables
Identificación	Variables necesarias para identificar los registros	ID-Cliente,Año-mes
Demográficas	Variables propias del cliente, sin importar su relación con la empresa	GSE, Edad, Genero, Potencial de gasto, Potencial de gasto 3 meses antes
De condición en la empresa	Variable de como la característica ve al cliente	Puntaje de riesgo, Puntaje de riesgo e meses antes
Variables de comportamiento	Variable de como el cliente compra con la tarjeta	Monto gastado en compras, Numero de rubros en los que compra, Monto Gastado hace 3 meses, Numero de rubros en los que compraba hace 3 meses, PGI con la tarjeta, PGI en rubros de (automotriz, combustible, comunicaciones, educación, entretención, farmacias , mejoramiento del hogar, recaudación, restaurantes, salud, seguros, supermercados, tiendas por departamento, transporte, viajes, vivienda),PGI en rubros de (automotriz, combustible, comunicaciones, educación, entretención, farmacias, mejoramiento del hogar, recaudación, restaurantes, salud, seguros, supermercados, tiendas por departamento, transporte, viajes, vivienda) hace 3 meses, numero de rubros en que compra fuera del holding, numero de rubros en los que compraba fuera del holding hace 3 meses, monto gastado fuera del holding, monto gastado fuera del holding hace 3 meses, activo en la tienda de retail del holding, activo en la tienda de mejoramiento del hogar del holding, deuda, numero de adicionales hace tres meses, numero de adicionales, enganchado con la tarjeta, enganchado con la tarjeta hace 3 meses
Variaciones en el comportamiento	Variables de como ha cambiado el comportamiento de los clientes en los últimos 3 meses	Variación porcentual en el potencial de gasto, variación porcentual en el monto gastado, variación porcentual en el PGI DE de la tarjeta, cambio en el numero de rubros, cambio en el numero de rubros on them, cambio en la frecuencia, cambio en el enganche
Tenencia de productos	Variables que muestran la tenencia de productos particulares	Puntos acumulados, porcentaje de canje de puntos, puntos vencidos, recency de puntos, tiene avance, tiene superavance, cuanto dinero tiene disponible de avance, cuanto dinero tiene disponible en superavance
Comunicación	Variables que muestran interés por comunicación/información	Numero de llamadas call center, activo vía web, numero de reclamos

Ilustración 3: Variables a considerar dentro de la base analítica con información de los clientes
Fuente: Elaboración propia con datos de la empresa

Esta base de datos contiene 21.837.424 registros y abarca datos desde marzo a octubre del año 2016.

Por su parte, la base de interacciones contiene las diferentes interacciones escritas que hacen los clientes tanto sobre servicios prestados por la empresa como sobre sus productos. La información recopilada en esta base proviene de tres fuentes: las solicitudes de servicios de los clientes, la tipificación de las interacciones habladas en el centro de llamados (Call Center) y una encuesta de satisfacción (conocida como NSP).

En los siguientes párrafos se describen con mayor profundidad las tres fuentes.

a) Solicitudes de Servicios: información manejada por el área de atención al cliente, incluye información de productos vendidos por la compañía o aquellos de los cuales el cliente deseó desvincularse.
b) NPS: Encuesta de satisfacción al cliente utilizada en la empresa. Utiliza una escala de números naturales (del 1 al 10), donde uno es la menor calificación y diez, la puntuación máxima, para medir el grado de satisfacción del cliente preguntándole por su disposición a recomendar los servicios prestados por la compañía. La métrica en este case se obtiene por medio de la resta entre el porcentaje de promotores (clientes con nota superior a 8) y el porcentaje de detractores (clientes con nota inferior a 7). Dentro de esta encuesta también existen comentarios escritos asociados a la nota, que justifican la nota en base a elementos de la compañía.
c) Call center: Esta base corresponde a una tipificación del contenido de las llamadas de diferentes clientes con el call center de la compañía. Esta llamada podría abordar diferentes tópicos como reclamos, preguntas, etc.

Estas 3 bases fueron unificadas en una tabla que posee los siguientes campos:

Variable	Descripción Variable
ID-Interacción	Indicador del registro dentro de la base
ID-Cliente	Indicador del cliente
Año-Mes	Variable que representa el año y mes en que se está estudiando las variables para un cliente determinado
Num_NPS	Numero de encuestas NPS respondidas por el cliente
Num_CC	Numero de llamadas al call center de la compañía por el cliente
Num_SS	Numero de solicitudes de servicios del cliente
Fecha_comentario	Fecha en la que se realiza el comentario
Texto	Texto de la interacción
Fuente	Indicador de la fuente de la que proviene el comentario

Ilustración 4: Variables a considerar dentro de la base analítica con interacciones escritas
Fuente: Elaboración propia con datos de la empresa

Finalmente, es importante destacar que esta base de interacciones contiene 5.508.442 registros, el primero de los cuales fue generado en noviembre del año 2003.

2.2 Indicación del Problema, Posibles Causas y Efectos

Para estudiar el efecto que tiene la fuga de clientes en la compañía, es necesario analizar previamente el comportamiento de un cliente tradicional. Para esto se utilizara la una marca de fuga generada dentro de la empresa que considera a los clientes fugados de la compañía como aquellos clientes que no realizan transacciones con la tarjeta dentro de los últimos 6 meses.

Año-mes	Numero de clientes	Numero de clientes fugados	Monto de gasto promedio por cliente ($ CLP)	Porcentaje fugado
ene-16	2.997.002	169.782	115.559	5,7%
feb-16	3.001.133	171.047	116.394	5,7%
mar-16	3.006.061	162.904	125.938	5,4%
abr-16	3.009.130	162.456	127.468	5,4%
may-16	2.815.863	160.171	131.067	5,7%
jun-16	2.821.129	190.461	123.963	6,8%
jul-16	2.826.315	187.881	129.507	6,6%
ago-16	2.832.483	188.781	127.845	6,7%
sep-16	2.835.735	196.811	122.094	6,9%
oct-16	2.843.888	199.025	130.500	7,0%
nov-16	2.853.044	205.009	136.623	7,2%
dic-16	2.871.347	178.485	167.973	6,2%

Tabla 1: Evolución clientes normales en el año 2016
Fuente: Elaboración propia con datos de la compañía

Si bien la definición de estado de fuga entregado por la compañía en base a la recencia es útil para analizar la evolución de los clientes, no es suficiente en este caso, dado que no considera la frecuencia de las compras realizadas con la tarjeta, por lo cual, no logra discriminar entre clientes que compran entre intervalos de tiempos pequeños (clientes de ciclo corto) y clientes que compran entre intervalos de tiempo más grandes (clientes de ciclo largo).Lo anterior sugiere la necesidad de determinar el criterio de fuga que mejor se adecúe a la realidad de la empresa estudiada.

Una posible causa de fuga de clientes se relaciona con la falla de aspectos específicos de la compañía. Algunos de ellos pudieran pasar inadvertidos para la empresa, mas ser de vital importancia para los clientes. Una manera de acercarse a las necesidades de los clientes es recurriendo a los comentarios escritos emitidos por ellos. El conocer aquello que los clientes piensan que debe mejorar al interior de la compañía, y subsecuentemente implementar medidas tendientes a la corrección de este punto, podría implicar retener a ese cliente, que aumentará su grado de satisfacción con la empresa.

La siguiente tabla presenta la evolución de los clientes que emitieron comentarios escritos durante 2016.

Año-mes	Numero de clientes	Numero de clientes fugados	Monto de gasto promedio por cliente ($ CLP)	Porcentaje fugado
ene-16	21.778	415	271.383	1,9%
feb-16	19.280	394	287.360	2,0%
mar-16	22.227	380	311.190	1,7%
abr-16	25.867	482	307.548	1,9%
may-16	26.706	440	308.604	1,6%
jun-16	31.378	653	270.940	2,1%
jul-16	38.495	845	257.990	2,2%
ago-16	41.743	1.067	265.451	2,6%
sep-16	33.539	769	279.553	2,3%
oct-16	32.249	745	305.422	2,3%
nov-16	33.463	732	339.945	2,2%
dic-16	30.415	597	381.734	2,0%

Tabla 2: Evolución clientes con comentarios escritos en el año 2016
Fuente: Elaboración propia con datos de la compañía

Al comparar los porcentajes de fuga de estos clientes con los presentados en la tabla 1 (que corresponden a clientes "normales"), teniendo en cuenta que ambas muestras corresponden al mismo periodo, se observa que quienes realizaron comentarios escritos presentan menores porcentajes de fuga. Más aún, los clientes que han realizado algún comentario escrito tienen una tasa de fuga tres veces menor que quienes no han realizado comentarios. Es importante considerar al mismo tiempo que estos clientes con comentario gastan aproximadamente dos veces más que los clientes sin comentarios.

2.3 Propuesta de Valor a la Empresa

A continuación, se presenta la evolución del valor total perdido en la cartera. Este se define como la evolución durante el periodo estudiado del dinero que se fuga de la compañía tanto para clientes con y sin comentarios.

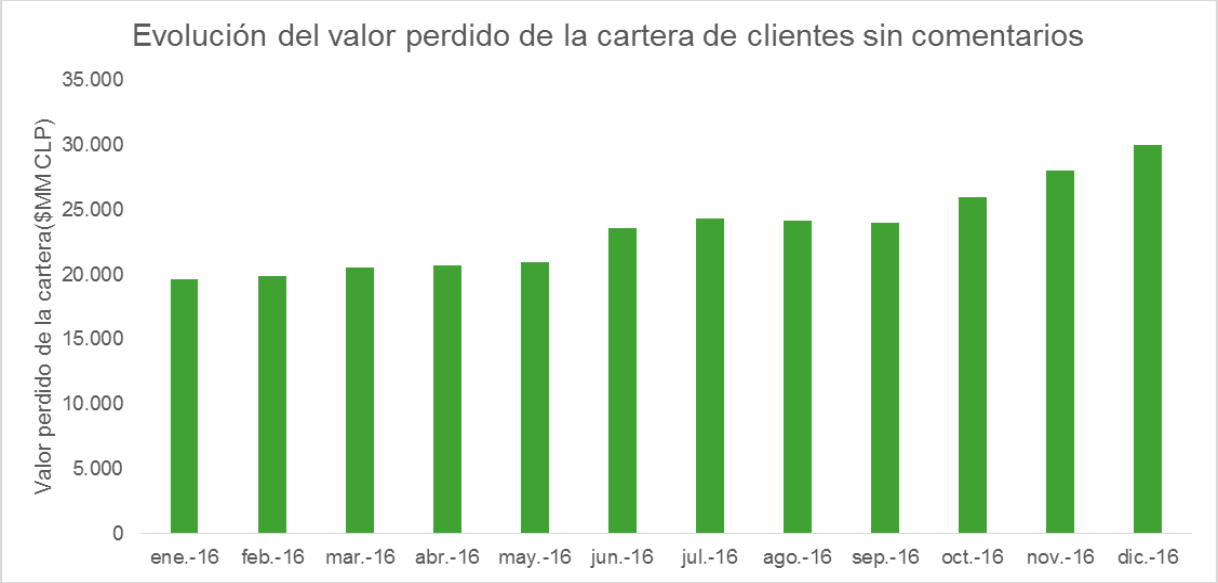


Grafico 8: Evolución del valor perdido en la cartera para clientes sin comentario
Fuente: Elaboración propia con datos de la compañía

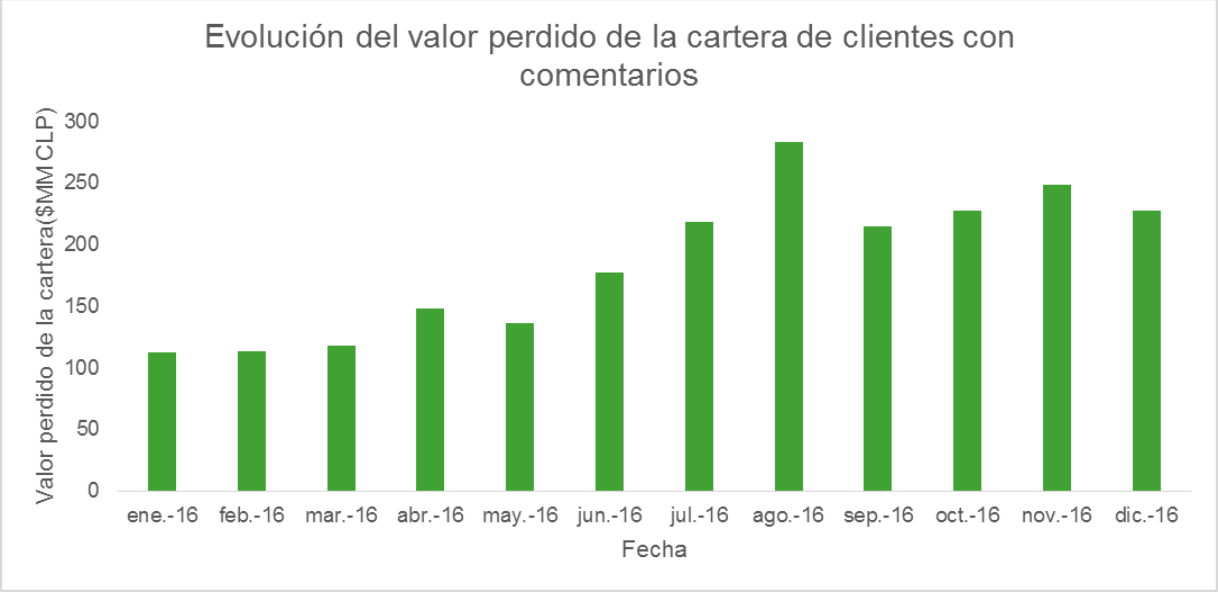


Grafico 9: Evolución del valor perdido en la cartera para clientes con comentario
Fuente: Elaboración propia con datos de la compañía

Se puede observar en los gráficos anteriores los clientes sin comentarios presentan un valor mayor en su totalidad que los clientes que si los tienen. Esto se debe principalmente a que, como fue presentado en las tablas anteriores, es una cantidad de clientes es considerablemente mayor de aquellos que no posee comentarios. Sin embargo, es importante considerar a los clientes con comentarios dado que estos textos escritos pueden dar luces más claras de los problemas que existen dentro de la compañía, pudiéndose con esto recuperar una mayor cantidad

de ese valor de la cartera tanto para clientes con y sin comentario. Es importante también considerar que el porcentaje de la cartera que se puede recuperar dependería de los resultados de la memoria, dado que se determinaría si los tópicos de comentarios son extrapolables a los clientes que no tienen comentarios.

La propuesta de valor para la empresa radicaría en lograr recuperar un porcentaje de estas pérdidas, tanto de los clientes con y sin comentarios. Esto se realizaría por medio de la extracción de información de los comentarios. La meta es procesar las percepciones de los clientes respecto a los aspectos de la empresa que consideran más relevantes, y el empleo de esta información dentro del modelo de fuga.

Para poder valorizar monetariamente lo que esta memoria representaría para la compañía se tomaran los datos del año 2016. En base a éstos se calcula que la diferencia porcentual entre los dos grupos es de aproximadamente 4,2%. Si se pudiera, por medio de los resultados generados por la memoria, reducir el porcentaje de fuga en un 0,1%, se generaría que aproximadamente 2892 clientes normales no se fugaran. Considerando esto último con que el monto gastado en promedio por un cliente normal durante un mes es de $129.578 pesos se tiene que el valor generado por esta memoria seria de 374.837.177 mensual en rentabilidad.

3. OBJETIVOS

3.1 Objetivo General

Construir un modelo predictivo de fuga utilizando textos para poder mejorar la retención de clientes.

3.2 Objetivos Específicos

1. Clasificación de los textos provenientes de diferentes puntos de contacto entre clientes y la empresa.

2. Determinar el impacto de las interacciones en el comportamiento transaccional de los clientes.

3. Construcción de modelos de propensión en base a variables demográficas, transaccionales, y de interacciones escritas de clientes.

4. Definir líneas de acción en base a los resultados obtenidos.

4. ALCANCES

Para el análisis de esta memoria se utilizan 3 canales por los cuales se podría recibir texto para evaluar el modelo. Estos son: Solicitudes de Servicios, Net Promoting Score (NPS) y del dentro de llamadas de la compañía (Call Center).

Sólo se considerarán clientes cuyos comentarios escritos puedan ser identificados de manera única y con fecha clara de realización de comentario. Esto se debe a que por medio de estos criterios se puede determinar al cliente estudiado y el cambio en su comportamiento desde el momento de la interacción con la compañía. Estas interacciones además deben de tener un largo superior a una palabra, dado que esto permitiría sacar conclusiones de cómo está siendo el servicio y con respecto a que se está hablando. Dados estos alcances, no se consideraran elementos de redes sociales, por la dificultad de asociar un único correo a un cliente.

Es importante destacar adicionalmente que solo se utilizaran datos históricos de la empresa para esta memoria, y que además no se consideraran experimentos de campo dentro de ella. De las cantidades de datos previamente mostrados, se utilizan 1.052.857 datos sin comentarios y 249.200 datos con comentario, lo que conforma una muestra total de 1.302.057 datos.

5. MARCO CONCEPTUAL

Durante esta sección se espera poder profundizar en aspectos metodológicos necesarios para el desarrollo de la memoria. Estos se separan principalmente en 3 grandes tópicos, presentados en la siguiente tabla.

Tópicos	Descripción
Minería de texto	En esta área de estudio se profundizan tanto herramientas necesarias para el procesamiento de datos de texto como modelos propios del área. Un ejemplo de esto es como lograr la extrapolación de tópicos y de análisis de sentimientos.
Homologación de clientes	Esta área explica el procedimiento para el calculo de la propensión a realizar un determinado tipo de interacción. Para esto se pasa a ver la herramienta de propensity score matching, y el cómo se puede llevar a cabo dentro de este contexto.
Arboles de decisión	Esto se estudia dentro del marco conceptual dado que el modelo de fuga está construido en base a un random forest, que es un tipo particular de árbol de decisión.

Ilustración 5: Temas del marco conceptual
Fuente: Elaboración propia

Por último se revisara elementos adicionales necesarios para entender el desarrollo de la memoria.

5.1 Minería de Textos

5.1.1 Definiciones Básicas

Esta sección pretende abordar los elementos básicos inherentes a la minería de texto, para poder así pasar a una descripción más completa de los algoritmos utilizados dentro del trabajo de título.

Una de estas definiciones es la de una palabra, que se define como un conjunto de caracteres que denotan un significado particular. Un conjunto de palabras generan elementos más complejos, llamados frases. La frase permite relacionar diferentes personas, lugares y acciones dentro de un determinado contexto. Un conjunto de frases se les llama documento, y cumple con la condición de cumplir un objetivo particular. [15]

Previo a la utilización de un modelo de minería de texto, es necesario la eliminación de palabras que aportan poco valor y del agrupamiento de palabras con el mismo significado. Las palabras que aportan poco valor dentro de este contexto se denominan stopwords, y son seleccionadas de acuerdo al lenguaje al que corresponde y al contexto en el que se utilizan las palabras. Para poder agrupar las palabras, estas se suelen llevar a su palabra raíz, para poder así entender mejor la relevancia que tiene entre documentos. Para poder hacer esto existen 2 metodologías principales para hacerlo, que son las de lematización y stemming. [12][14][16]

Por una parte la lematización toma las terminaciones de las palabras y al detectar si es un verbo o un sustantivo pasa a hacer un procesamiento de esta información tal que los sustantivos pasan a su versión singular y los verbos pasan infinitivo, para lograr hacer un agrupamiento. La técnica de stemming lo que hace es eliminar la terminación de las palabras, para poder así simplificar la palabra y que sea más fácil de agrupar. Una vez realizado alguno de estos procesamientos, es posible implementar los modelos de Retiro de información y de sentimiento de texto.

Dentro de este contexto, se decide utilizar las técnicas de lematización por sobre las de stemming, dada la facilidad de entender los significados de las palabras, y poder hacer gestión sobre ellas. Es además apoyada dentro de la memoria de Constanza Contreras por el mismo motivo [16]

5.1.2 Algoritmos de Informtion Retrieval (IR)

Los algoritmos de retiro de información (Information retieval) son los algoritmos responsables de resumir la información disponible en conjuntos de documentos para poder así determinar elementos en común y poder agruparlos. Algunos de estos toman una perspectiva probabilística para lograr determinar estos grupos como lo son el caso de latent dirichlet allocation model (LDA) y Pitman-yor Topic Model (PYTM).

LDA

LDA se basa en la existencia de k tópicos, de los cuales se generan los documentos que se tienen. Cada tópico es representado por una distribución multinomial por todas las palabras existentes en los documentos, que se define como el vocabulario. Un diccionario para un tópico se genera en base a la mescla entre las palabras del vocabulario y la probabilidad de pertenecer al tópico.

En un caso más concreto, un documento <w1, w2, w3,.., wn> es generado en base al siguiente proceso. Primero se toma θ que sigue una distribución de Dirchlet con parámetros $<\alpha_1, \alpha_2, \alpha_{3,..}, \alpha_k,>$ Con esto, un punto dentro del grafo donde θ_i>0 y $\sum \theta_i$=1. Luego, para cada una de las n palabras es tomada como muestra de la multinomial (θ) con lo que la probabilidad de pertenecer al tópico i está dada exclusivamente por θ, y que genera θ_i. [5][6][16]

Matemáticamente, esto se presenta como a continuación:

$$p(\theta,z,w|\alpha,\beta)=p(\theta|\alpha)\Pi p(z_n|\theta)p(w_n|z_n,\beta)$$

Ecuación 1: Calculo de las probabilidades para LDA

Donde:

- $\boldsymbol{w_n}$=palabra n-esima
- θ =distribución de tópicos
- α =parámetro de la distribución de Dirchlet
- β =matriz de apariciones de palabras y tópicos
- $\boldsymbol{z_n}$ = tópico asociado a la n-esima palabra

PYTM

PYTM es un algoritmo basado en LDA donde la distribución que la muestra theta tomaría no estaría restringida a una distribución de Dirchlet como es el caso de LDA, y esto lo realiza por medio de proceso de restaurante chino.

Proceso del restaurante chino

Este proceso argumenta que si existe n clientes, y tienen infinitas mesas para sentarse, la probabilidad de que se siente en la mesa t es

$$\begin{Bmatrix} \frac{|bt|}{n+1} & Probabilidad\ de\ sentarse\ en\ una\ mesa\ ocupada \\ \frac{1}{n+1} & Probabilidad\ de\ sentarse\ en\ una\ nueva\ mesa \end{Bmatrix}$$

Ecuación 2: Calculo de la probabilidad de sentarse en la mesa t

Donde |bt| es el número de clientes en la mesa t

El proceso de restaurante chino se modifica para poder generar el algoritmo de Pitman Yor. Si imaginamos que el número de personas aumenta el número de mesas ocupadas también lo hará. Modificando la ecuación 2 se puede llegar a la siguiente expresión:

$$\begin{Bmatrix} \frac{N_{j,k}^{c}-d}{\gamma+N_{j}^{c}} & Probabilidad\ de\ que\ la\ mesa\ k\ este\ ocupada \\ \frac{\gamma-dK_{j}}{\gamma+N_{j}^{c}} & Una\ nueva\ mesa\ desocupada \end{Bmatrix}$$

Ecuación 3: Cálculo de la probabilidad de sentarse en la mesa t cuando existe un número finito de tópicos

Donde $N_{j,k}^{c}$ es el número de consumidores sentados en la mesa j, $N_{j}^{c} = \sum_{k} N_{j,k}^{c}$ indica el largo del documento y K_j indica el número de mesas en el restaurante j.

Con esto se logra modificar la distribución dada las palabras en cada tópico hasta el momento, que puede ser modificado para tener un proceso generativo que genera el algoritmo de Pitman Yor. [16]

5.1.3 Análisis de Sentimientos

Para lograr asociar un sentimiento a un texto determinado la literatura recomienda 2 caminos principales, uno por medio de diccionarios lexicográficos y otro por medio de modelos construidos en base a machine learning.

Técnica para determinar el sentimiento	Descripción
Sentimiento construido por herramientas de machine learning	Este mecanismo genera una cierta ponderación a cada palabra presentada en un documento para poder determinar si el texto es positivo o negativo. Para lograr desarrollar esto es necesario tener textos etiquetados para lograr calibrar el modelo.
Sentimiento construido por medio de diccionarios léxico-gráficos	Este mecanismo toma una serie de diccionarios que tienen palabras con una determinada asignación de polaridad. Por medio de la ponderación un puntaje en base a estas palabras es posible obtener los resultados de polaridad.

Ilustración 6: Técnicas de obtención de sentimientos
Fuente: Elaboración propia

Para este estudio se decide utilizar un diccionario léxico-gráfico, dado los resultados presentados en la memoria de Cortez [12].

El modelo de esta memoria requiere de un diccionario con palabras positivas otro con palabras negativas, que además deben poseer un determinado grado de polaridad que tienen las palabras a considerar. En el caso de este estudio, se toman los niveles de alto medio y bajo. Estos niveles y polaridades son luego transformados a un valor numérico en base a si cuan alto es su nivel (si el nivel de una palabra es alto se le asigna un puntaje de 3, si es medio un puntaje de 2, y si es bajo un puntaje de 1) y a su polaridad (si la palabra es una palabra negativa el valor es multiplicado por -1). Luego de la valorización de las palabras, se pasa a la normalización del valor, por medio de la suma de todas las palabras en cada diccionario y posterior división por el total. Por otra parte existe un tercer diccionario de palabras intensificadoras. Estas palabras son aquellas que afectan a la palabra que les sigue. Por ejemplo "no" es una palabra que logra invertir la polaridad de las palabras conceptualmente. Para que pueda hacer esto matemáticamente debe de multiplicar los valores por -1. Este diccionario será considerado en el futuro.

Para poder determinar el efecto de una palabra es necesario encontrar primero al sujeto de la oración. Esto se hace en este caso por la construcción de un diccionario de posibles sujetos relacionados a la empresa. Luego de esto se inicializa un puntaje negativo y positivo, y a medida que se recorre el texto se le va sumando al puntaje el valor presentado en la siguiente ecuación.

$$Puntaje = Puntaje + \alpha * D(p) * e^{\frac{-pos^2}{2}}$$

Ecuación 4: Calculo puntajes sentimientos con diccionario léxico-grafico

Donde:

- α es el valor del intensificador que estaba previo a la palabra, en caso de existir
- D(p) es el valor de la palabra dentro del diccionario del palabras positivas o negativas
- Pos es la distancia entre la palabra estudiada y el sujeto de la oración

Una vez que se ha desarrollado esto, se puede pasar a la evaluación final del texto, en donde la fórmula para asignar puntaje es la siguiente

$$puntaje\ texto = \varphi(Puntaje\ positivo\ + Puntaje\ negativo\ )$$

Ecuación 5: Calculo puntaje total de un comentario con diccionario léxico-grafico

Donde el φ es un multiplicador de expresividad. Si el texto contiene un signo de exclamación o una repetición sucesiva de muchas vocales pasa a ser un valor de 1.3. En caso contrario, se mantiene en un valor de 1.

5.1.4 Distancias

Una de las herramientas utilizadas para lograr comprobar la similitud entre diferentes textos es la distancia. Estas distancias son métricas que permiten indicar cuan similares son documentos (si se estudia la distancia entre palabras) o palabras (si se estudia la distancia entre caracteres).

Un ejemplo de distancia consiste en la distancia coseno. Esta distancia, por medio de la descomposición en planos vectoriales de palabras, un documento queda definido por un único vector. Calculando la distancia coseno entre ambos ángulos es posible tener una métrica de distancia que es 0 de ser dos elementos idénticos o 1 si no tienen elementos en común. [16]

5.2 Modelos para la Construcción de Fuga

Para lograr hacer la homologación de clientes que tienen comentarios con aquellos que no los tienen se utiliza la técnica de propensity score matching, que es una técnica cuasi-experimental que permite comparar clientes que han tomado un tratamiento (en este caso, la realización de uno de los comentarios) con aquellos que no lo han tomado. [10][8]

Esto se realiza por medio de los siguientes pasos:

5.2.1 Estimación de un Modelo de Propensión

Este paso consiste en la toma de variables de los clientes previos al tratamiento y pretende en base a esto estimar la probabilidad de que tomara uno de los tratamientos. En este caso, las variables previas al tratamiento abordan características que podrían afectar a la realización de un determinado comentario. En este caso se tomaron las siguientes variables.

Variable Logit multi-variado
Edad
Sexo
Tenencia Sav
PGI Tarjeta
Puntos vencidos
Recency Puntos
Cliente tiendas por departamento
Cliente tiendas de mejoramiento del hogar
Cliente de uso regular de la tarjeta
Deuda
Tipo de tarjeta
Sexo
GSE
Disponible Súper avance
Disponible avance

Ilustración 7: Variables a considerar dentro del logit multivariado
Fuente: Elaboración propia con datos de la compañía

Para poder hacer el cálculo de la propensión se utilizó un logit multivariado. Dentro de las posibilidades de realizar un modelo de propensión multivariado, la alternativa de un logit multivariado presentaba una simplicidad computacional mayor al probit multivariado y contaba con la exactitud que no presentaba la posibilidad de realizar una serie de logits bi-variados múltiples. [8]

Para lograr comprender el funcionamiento del logit, pasaremos a revisar su definición

Logit (Bi-variado)

Este modelo se desprende de la suposición de que la valoración que se le da a un determinado elemento está dado por la una componente visible, y un factor aleatorio de error, que genera que esta variabilidad se vea afectada para distintas personas

$$v = u + \varepsilon$$

Ecuación 6: Supuesto de la utilidad de una decisión para el logit

Donde v representa la utilidad para un determinado cliente, u la utilidad en base a los atributos de un determinado producto y de la componente aleatoria de la utilidad de distintos clientes.

Lo que la teoría sostiene es que por medio de la comparación de alternativas uno puede calcular la diferencia entre los distintos aspectos existentes entre las diferentes alternativas. Cuando uno calcula la probabilidad de esto, el cálculo toma la siguiente forma:

$$P(v_{ni} > v_{nj}) = \int 1_{[\varepsilon_{nj} - \varepsilon_{ni} < v_{nj} - v_{ni}]} f(\varepsilon_n) d\varepsilon_n$$

Ecuación 7: Probabilidad de selección de una alternativa por sobre la otra

Asumiendo que la distribución del error tiene la forma de una doble exponencial, se llega a la forma tradicional del logit bi-variado.

$$P_{ni} = \frac{e^{u_{ni}}}{\sum e^{u_{nj}}}$$

Ecuación 8: Formula cerrada del cálculo del logit bi-variado

Para lograr pasar a la fórmula del caso multivariado, se toma en vez una distribución del error que sigue una multinomial. Este reemplazo dentro de la integral presentada lleva a la fórmula del logit multivariado.

5.2.2 Elección del Algoritmo de Matching para encontrar los Clientes Homólogos

El algoritmo de matching identifica a los clientes que presentan propensiones más próximas entre ellos. Esto se puede realizar por medio de:

a) KNN: Identificación de los K vecinos más cercanos a un determinado dato dentro de los planos en que se encuentre.
b) Caliper y Radius: Considera a los vecinos más cercanos dentro de un radio definido, pero al mismo tiempo establece una distancia mínima que se debe cumplir para indicar que estos vecinos sean considerados.

Durante esta memoria se utiliza el método de KNN, donde se consideran los 3 vecinos/as cercanos para el estudio del comportamiento transaccional, y al más cercano para el estudio de los clientes homologados dentro de la fuga. [8][10]

5.3 Modelos basados en Árboles

5.3.1 Information Gain

Information Gain es una definición importante dentro de esta memoria dado que se utiliza para eliminar outliers y como criterio de división dentro de los árboles utilizados en la memoria. Este criterio presenta la siguiente fórmula. [10][13][16]

$$Information\ Gain\ (S, F) = Entropia(S) - \sum_{v\ \in Valores(F)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Ecuación 9: Ecuación de Information Gain

Donde S_v es un subconjunto de datos con el valor v de la variable F, y se logra determinar cual es el mejor valor para realizar el corte. Es importante destacar al mismo tiempo que la entropía es una medida de dispersión utilizada dentro de esta ecuación, que se calcula de la siguiente forma.

$$Entropia(S) = \sum_{c\ \in Valores(S)} -\frac{|S_c|}{|S|} log_2 \frac{|S_c|}{|S|}$$

Ecuación 10: Ecuación de Entropía

5.3.2 Arboles de Decisión

Los árboles de decisión consisten en herramientas matemáticas construidas para lograr hacer clasificación y regresión de variables independientes por medio de la división de los diferentes grupos de datos en elementos homogéneos. Esto se utilizará en el contexto de esta memoria para lograr homologar clientes por medio de variables similares, como se explicara en la sección de metodología.

El principal tipo de árbol que se construyó para el desarrollo de esta memoria fue el árbol de tipo CART. Esto se debe a que es el tipo de árbol que se utiliza dentro del modelo de random forest seleccionado.

Los árboles CART utilizan como objetivo una variable categórica, y hace divisiones binarias sobre los datos disponibles. Esto quiere decir que un nodo padre presenta siempre 2 hijos.

Para lograr comprender bien como se realiza la división, es necesario entender el criterio de Information Gain, que fue presentado anteriormente. [10][13]

5.3.3 Random Forest

Este modelo predictivo se basa en la construcción de varios árboles de decisión, por medio de diferentes combinaciones lineales de las variables dependientes para poder así predecir la variable dependiente. Al tener la construcción de los diferentes árboles se selecciona el árbol que tenga un mayor poder predictivo dentro de los arboles construidos por el algoritmo. [9]

El random forest en particular que se utilizará durante esta memoria es el presente en el programa de SPSS modeler (Versión 18). Este construye sus árboles en base a arboles CRT, que fueron presentados anteriormente.

Uno de las desventajas que presenta el random forest como elemento de análisis para estudios en más profundidad es que es un algoritmo que simplemente entrega un resultado sin saber necesariamente cuál de los árboles es el que entrega el resultado, haciendo imposible la evaluación de los árboles individuales de los que se obtienen los resultados.

Un punto importante a considerar para este modelo es como se explican las variables importantes dentro de este modelo. Para lograr determinar la importancia de cada uno de los predictores se itera la construcción del modelo sin esa variable en particular, y se compara la precisión del modelo cuando esta variable es incluida.

5.4 Otros Modelos

5.4.1 Estudio de Recency, Frecuency y Monetary Value (RFM)

Metodología que permite describir el comportamiento del uso de un determinado cliente de algún producto y/o servicio provisto por alguna compañía. El RFM se compone de:

- Recency: número de días desde la última utilización de la tarjeta
- Frequency: número de usos de la tarjeta dentro de un determinado intervalo de tiempo
- Monetary Value: monto gastado por parte del cliente dentro de un intervalo de tiempo

Esta metodología es comúnmente utilizada para lograr definir la fuga de un cliente, y en base a esto construir el modelo de fuga. Es importante considerar además, que en algunos casos estos indicadores se suele usar de manera independiente, como es el caso actual de la compañía donde se usa exclusivamente el recency. Como se mencionó anteriormente la combinación de diferentes elementos puede reflejar de mejor manera la realidad. Uno de estos casos es el

indicador de recency dividido en frecuency, que indica el tiempo entre compras proporcional a la frecuencia de compra de un determinado cliente. [1][2][3][4][14]

6. METODOLOGÍA

La metodología a utilizar dentro de esta memoria consistiría en la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [9]. Como el título de esta metodología sugiere, es un proceso que pretende estandarizar los pasos para el desarrollo de proyectos de minería de datos, y esto lo realiza por medio de 6 pasos:

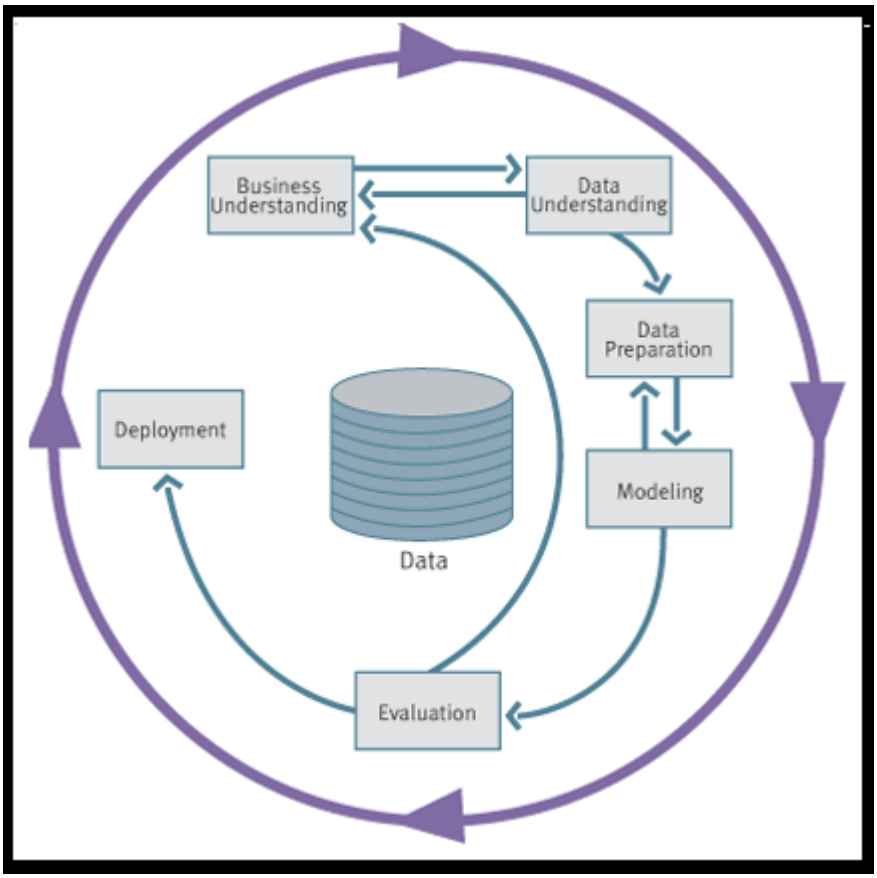


Ilustración 8: Proceso CRISP-DM

6.1 Entendimiento del negocio

Esta etapa se define los objetivos de un determinado proyecto. Además se consideran tanto los recursos que se tienen como los que son necesarios.

6.2 Entendimiento de la data

Esta etapa se caracteriza por el estudio de la data con la que se cuenta. Además de esto se realiza un estudio de la calidad de esta data, para poder proteger la precisión del modelo.

Durante esta etapa se revisó la data disponible, cuáles eran sus fuentes de origen y de cómo se relacionaban entre ellos. Además de esto, se realizó un estudio univariado para lograr determinar cuáles de las variables se relacionaban más fuertemente con la fuga.

6.3 Preparación de la data

Durante esta etapa se generan todas las transformaciones necesarias para generar los modelos dentro de la etapa de modelado.

Para iniciar el procesamiento de la data fue necesario construir una base analítica, a partir de la cual se pudieran construir los modelos necesarios. Esto se abordó por medio de 2 bases analíticas, una de las interacciones realizadas y otra con los datos de los clientes. Esto se tomó para el periodo ya especificado en las secciones previas.

Una vez logrado esto, se pasó a la construcción de todas las variables que debían ser generadas para introducir al modelo de fuga. Esto apunta por una parte a las variables generadas por medio de minería de texto y el modelo de la homologación de clientes.

Dentro de lo desarrollado para la parte de minería de texto fue necesario el pre-procesamiento de los datos. Esto incluye la eliminación de stopwords y lematizacion de los diferentes textos. Posteriormente, se pasó a la construcción de los modelos de Pitman Yor y de análisis de sentimiento.

Dentro de lo desarrollado para la homologación de clientes, se construyó el modelo de propensión por tópico y la homologación de clientes.

Por último, esta sección concluyó con la eliminación de outliers dentro de los datos y de una posterior selección de variables. Esto último se realizó por medio de los criterios de correlación y de information gain.

6.4 Modelado

En esta etapa se utilizan los datos obtenidos de las etapas previas para lograr generar modelos que aborden los objetivos planteados en la etapa inicial de entendimiento del negocio.

Durante esta etapa se lograron construir 4 modelos de fuga, para explicar diferentes aspectos de la fuga de clientes. Uno de ellos fue un modelo de fuga tradicional. Este fue desarrollado para ser usado como referencia de los otros modelos y poder realizar comparaciones. Luego, se construyeron dos modelos con variables de minería de texto, para observar cuál es el efecto que tiene en la fuga de clientes la idea de realizar un cometario. Por último, se construyó un modelo de clientes homologados determinar si era posible traspasar los conocimientos de una persona que realiza interacciones a una persona que no los realice.

6.5 Evaluación

En esta etapa se evalúa si los resultados obtenidos logran cumplir los objetivos planeados al inicio del proyecto. Una vez completada esta etapa, se determinan posibles líneas de acción dados los resultados obtenidos por parte de los modelos.

Dentro de esta sección, se analizaron los resultados obtenidos y se seleccionó el mejor de fuga para esta compañía de entre los modelos disponibles. Posterior a esto, se realizó un análisis de sensibilidad con respecto a las variables disponibles, principalmente para poder revisar como varían los errores de tipo I y tipo II.

6.6 Ejecución

En esta etapa, se planean, ejecutan y monitorean los diferentes proyectos. Con esto se completa el proyecto desde la perspectiva de minería de datos. Dada la naturaleza del trabajo de título se pretende llegar a la etapa de evaluación del proyecto (etapa previa), no obstante se buscó dar líneas de acción de cómo seguir con respecto al proyecto.

7. RESULTADOS OBTENIDOS

7.1 Definición de Fuga

Uno de los requisitos para la realización de un modelo de fuga consiste en poder definir de manera adecuada el criterio de ésta.

La fuga estudiada sería la fuga no contractual de la tarjeta. Ésta se define como la fuga relacionada a un cliente que deja de hacer uso de su tarjeta, si bien puede seguir teniendo un contrato con la compañía. Para poder definirla es importante dentro de la empresa que éste criterio de fuga cumpla con las siguientes características:

- Sostenible en el tiempo: Este aspecto se refiere a que los clientes considerados como fugados se mantengan en este estado por la mayor cantidad de tiempo posible, y que al mismo tiempo logre minimizar la reactivación de los clientes de la compañía. Esto permite que la misma métrica sea utilizada en el futuro sin considerar mayores variaciones con respecto a lo visto anteriormente.

- Tome una masa crítica de clientes que se asocie a la realidad de la compañía: Este criterio apunta a que la definición de fuga logre capturar una cantidad de clientes que sea proporcional a actividad que tiene la compañía. Esto quiere decir que se tome una masa de clientes que sea lo suficientemente grande para poder representar a los clientes fugados, pero que no sea tan grande y que la mayoría de los clientes sean fugados.

Para definir la fuga se recurrió a los indicadores de RFM, de manera tal de observar la variación del comportamiento del cliente a lo largo de un año tradicional. A continuación se presentan los gráficos para R, F, M y R/F. Para abordar el análisis, se decide tomar una muestra del 5 % de los clientes de la cartera para el año 2016. Con esto se pretenden utilizar 3 métricas para comparar los diferentes indicadores:

- Variación del indicador dentro de un mismo periodo de tiempo
- Variación del indicado con respecto al periodo de tiempo anterior
- Análisis de reactivación de clientes

Los resultados de cada una de estas secciones se muestran a continuación.

7.1.1 Variación del indicador dentro de un mismo periodo de tiempo

Previo al estudio de la variación, es útil hacer una revisión de cómo se comportan cada uno de los indicadores en los meses del año 2016. Éstos se presentan a continuación.

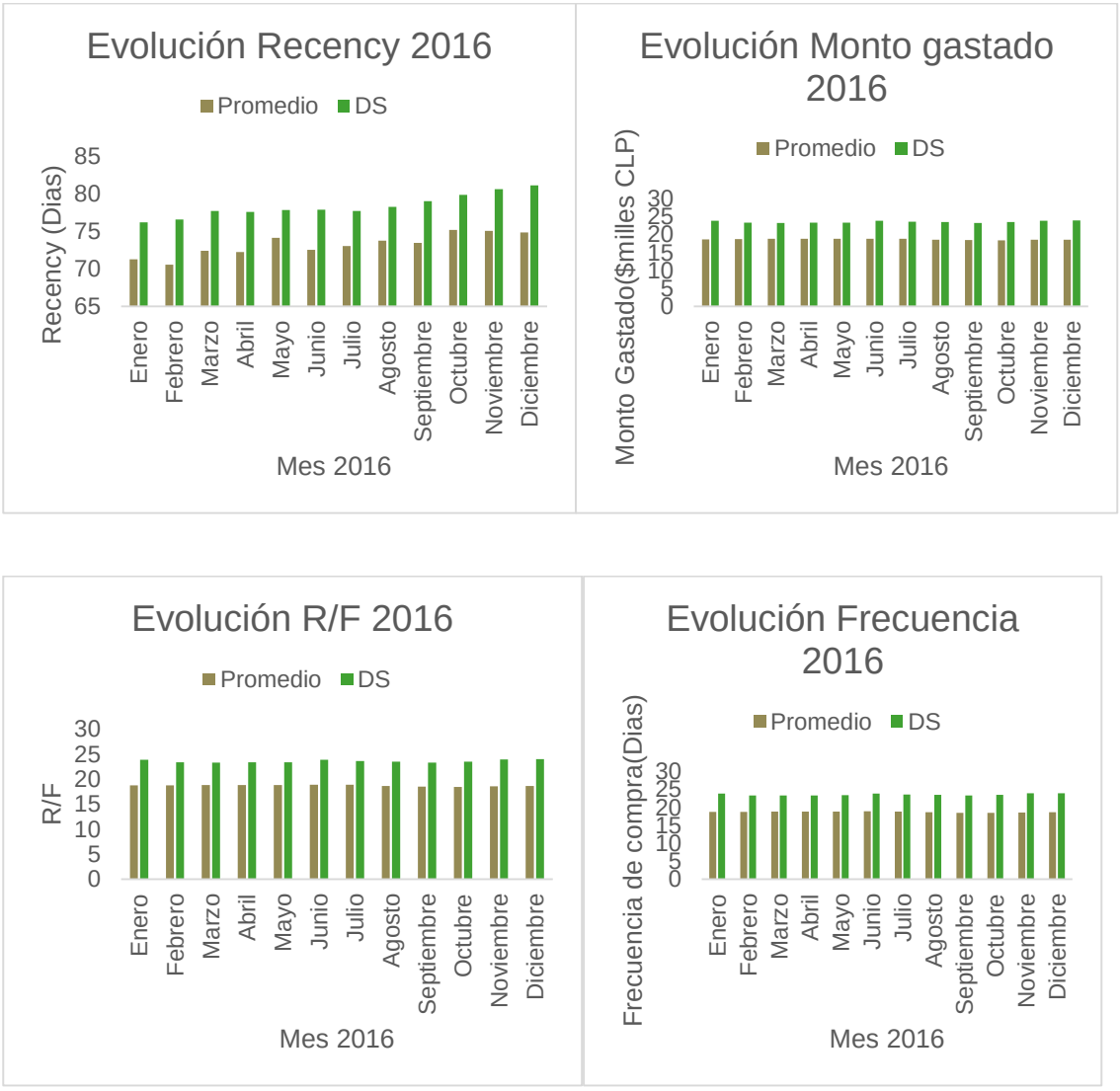


Grafico 10: Gráficos evolución de las diferentes métricas estudiadas
Fuente: Elaboración propia con datos de la compañía

Se puede apreciar que dentro de las métricas presentadas la que presenta una mayor variación con respecto a su promedio es la recencia de la compra. Los otros indicadores se presentan como criterios más estables en el periodo estudiado.

Para poder resumir el grado de dispersión de los datos, se realizó un gráfico para representar cómo varía cada métrica en un determinado mes. Además, se determinó la distancia intercuartil promedio para cada una de las métricas, las cuales se presentan a continuación.

Métrica	Distancia Intercuartil
Recency	0,35
Monto Gato	0,66
R/F	0,00
Frecuencia	0,82

Tabla 3: Distancia Intercuartil de cada una de las métricas para un periodo de tiempo
Fuente: Elaboración propia con datos de la compañía

Como se puede apreciar en la tabla, la métrica que presenta la mayor dispersión es el monto gastado por los clientes, por el contrario, aquella que presenta menores variaciones dentro de un determinado mes corresponde a la métrica de R/F.

7.1.2 Variación del indicador con respecto al periodo de tiempo anterior

Para lograr abordar las variaciones entre diferentes periodos de tiempo, se toman las variaciones entre meses para cada una de las métricas. En base a esta variación se realiza un boxplot y se calcula la distancia intercuartil de cada una de las métricas, lo que se presenta a continuación.

Métrica	Distancia Intercuartil
Recency	0,99
Monto Gato	0,61
R/F	0,75
Frecuencia	0,25

Tabla 4: Distancia Intercuartil de cada una de las métricas para un periodo de tiempo
Fuente: Elaboración propia con datos de la compañía

Con respecto a este estudio se puede ver que la recencia presenta el peor de los desempeños. Por otra parte, la métrica de frecuencia es la más estable en este sentido.

7.1.3 Análisis de reactivación de clientes

La última métrica a revisar consiste en lograr determinar cómo se considera la reactivación de clientes dentro de los meses sucesivos. Para esto, se tomó una muestra aleatoria de clientes y se estudió como es su comportamiento en los meses sucesivos a la fuga, para poder así ver el porcentaje de clientes que se vuelve a reactivar dentro de esos periodos. Para poder hacer esto, se realizaron una serie de cortes para todas las métricas, de manera de que se tuvieran cantidades relativamente homogéneas de fuga inicial, para así poder comparar las tasas de reactivación luego de 3 meses de lo ocurrido. Estos árboles completos se presentan en el Anexo I, II y III.

El resumen del análisis de los árboles se presenta a continuación

Métrica	3 meses	6 meses
R/F	0,46	0,50
Recency	0,50	0,55
Frecuencia	0,37	0,44

Tabla 5: Porcentaje de reactivados de la muestra en los próximos 3 y 6 meses
Fuente: Elaboración propia con datos de la compañía

Métrica	3 meses	6 meses
R/F	0,10	0,06
Recency	0,10	0,05
Frecuencia	0,05	0,03

Tabla 6: Porcentaje de fugados de la muestra en los próximos 3 y 6 meses
Fuente: Elaboración propia con datos de la compañía

Se puede apreciar de la tabla que si bien la frecuencia y recencia de las compras son un mejor indicador en un periodo de tiempo corto, el R/F es un mejor indicador en el largo plazo. Esto puede deberse a que tanto la recencia como frecuencia establecen un corte de fuga que se usan de manera idéntica para clientes que tienen distinta temporalidad de compra. Es por esto que para clientes que compran con cierta periodicidad (como por ejemplo una vez cada 6 meses), estos criterios de fuga no aplican de buena manera como si lo hace el criterio de R/F.

Se puede apreciar de las tablas 5 y 6 que la métrica que tiene menores variaciones es la de R/F, tanto para para la fuga como para la reactivación

.

7.1.4 Definición de Fuga

Dado los resultados anteriores, se concluye que la mejor métrica a utilizar es la de R/F. Esto se debe a que tiene un mejor desempeño en 2 de las 3 medidas utilizadas, por lo que es la métrica a utilizar. Sin embargo, es importante destacar que para lograr evitar problemas de asignación en la fuga se pretende poner un recency mínimo. Esto se debe a que si por ejemplo una persona que compra todos los días deja de comparar por 2 meses, puede que sea clasificado como fugado cuando solo no estuvo activo por un tiempo particular.

Para poder determinar el punto de corte que se le dará a la fuga de la compañía, se construye la siguiente tabla.

		Dias de Recency						
		0	30	60	90	120	150	180
Valor R/F	0	100,0%	16,0%	10,8%	7,8%	5,8%	4,3%	3,2%
	1	40,9%	13,8%	10,1%	7,5%	5,7%	4,2%	3,2%
	2	16,7%	10,5%	8,6%	6,8%	5,3%	4,1%	3,1%
	3	9,8%	8,0%	7,2%	6,1%	4,9%	3,8%	3,0%
	4	7,1%	6,4%	6,0%	5,3%	4,5%	3,6%	2,8%
	5	5,6%	5,3%	5,0%	4,6%	4,1%	3,4%	2,7%

Tabla 7: Porcentaje de fugados de la muestra en los próximos 3 y 6 meses
Fuente: Elaboración propia con datos de la compañía

Dado que la fuga actual de la empresa se encuentra próxima al 7% se decide por un corte de R/F>3 y Recency>60 días. Aquellos clientes que cumplen con estos criterios serían considerados como fugados.

7.2 Minería de texto

Esta sección tiene como objetivo describir el procesamiento necesario de información para poder extraer de los textos disponibles diferente información, que funcionara como variables de entrada para lograr predecir la fuga de la compañía.

7.2.1 Pre-Procesamiento de los datos

Previo a poder hacer uso de los algoritmos descritos en la sección de metodología, fue necesario el uso de técnicas para poder extraer la información más relevante de los diferentes textos. Estas técnicas fueron las de eliminación de stopwords y de algoritmos de stemming. La

eliminación de stopwords es un paso importante para poder eliminar palabras que aportan poca información al texto, para poder así dejar solo los elementos importantes de estos. Por otra parte la lematizacion es un proceso necesario para poder agrupar las palabras que tienen significados similares. Ejemplos de cada uno de estos casos se presenta a continuación.

Texto	Texto lematizado	Texto con stemming
Buenos productos. Promociones de puntos más pesos que es conveniente cuando no alcanzan los puntos	conveniente alcanzar punto promoción bueno peso producto	convenient alcanz punt promocion buen pes product

Ilustración 9: Ejemplo de lematizacion y steming
Fuente: Elaboración propia con datos de la compañía

Para lograr llevar a cabo la minería de texto fue el lenguaje de programación Python. Esto se debe principalmente a la amplia variedad de paquetes que existen dentro del lenguaje para poder hacer minería de texto y minería de datos en general. Algunas de estas librerías son nlkt (minería de texto especifico, particularmente útil para la realización de stemming y la eliminación de stopwords), pattern.es (módulo de Python desarrollado por la universidad de antwerpen para estudio de lenguajes en diferentes idiomas) y vpvy (librería que implementa el modelo de pitman-yor).

Uno de los casos especiales con respecto al manejo de textos fue el caso de los datos del call center. Esto se debe principalmente a que estos datos contaban con la complejidad de ser datos tipificados por las personas que participan en el call, siendo estos escritos sin el espacio de separación de palabras. Para poder hacer frente a esta situación se utilizó un algoritmo de separación de texto. Este consistía en lograr tomar las palabras de mayor repetición dentro del diccionario de la real academia española, y en base a la posición dentro del ranking dividía los textos para poder así lograr separarlos. A continuación se presenta un ejemplo de esto.

Texto original	Texto corregido
DONGABRIELCLDEJALA CONSTANCIADEQUEEL DIADEHOYSELEENVIOA SUCORREOUNMENSAJ ECONUNCODIGODEVAL IDACIONPEROCLNOHA REALIZADONINGUNAGE STIONSECOMUNICOEL DIADEHOYPEROENELSI STEMANOFIGURANINGU NATRANSACCION.	don gabriel cl deja la constancia de que el dia de hoy se le en vio a su correo un mensaje con un codigo de validacion pero cl no ha realizado ninguna gestion se comunico el dia de hoy pero en el sistema no figura ninguna transaccion

Ilustración 10: Ejemplo resultado de algoritmo de separación de texto
Fuente: Elaboración propia con datos de la compañía

7.2.2 Análisis de sensibilidad de Pitman-Yor

Si bien la literatura recomienda entre 25 y 30 tópicos, para lograr hacer una comparación de la eficiencia de los tópicos obtenidos con el algoritmo, se decidió hacer un análisis de sensibilidad con respecto a los tópicos obtenidos. Esto se realizaba mediante la comparación de similitud entre las palabras obtenidas de los diferentes tópicos variando tanto la cantidad de tópicos como la cantidad de iteraciones que se utilizaban en el algoritmo.

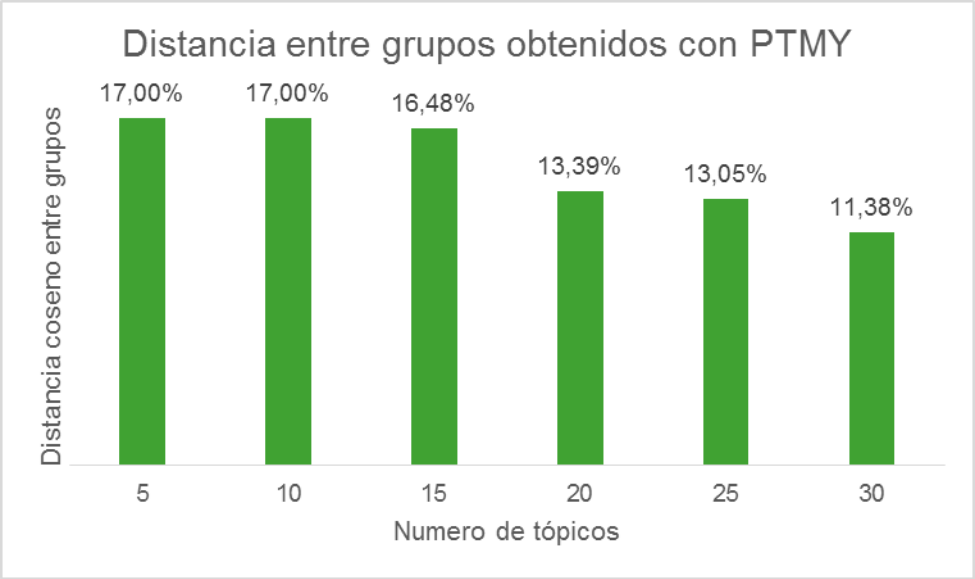


Grafico 11: Análisis de sensibilidad con el número de tópicos
Fuente: Elaboración propia con datos de la compañía

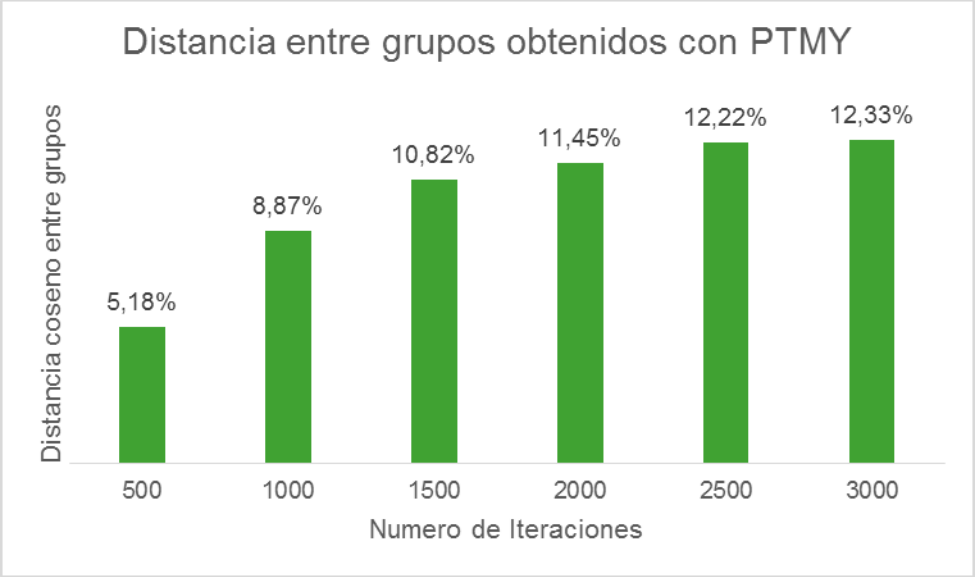


Grafico 12: Análisis de sensibilidad con el número de iteraciones
Fuente: Elaboración propia con datos de la compañía

Por una parte, se puede ver que la distancia entre los distintos tópicos empieza a disminuir una vez que se pasa de los 10 tópicos. Es por esto que, para poder hacer gestión con estos datos, se decide por esta cantidad de tópicos. Ocurre algo similar con el número de iteraciones, con lo que se decide por 2.500 dado el aporte marginal de más iteraciones no varía de manera sustancial la distancia obtenida.

7.2.3 Resultados minería de texto

Con respecto a los resultados obtenidos de la implementación del modelo de Pitman-Yor, se pueden ver las 10 palabras más influyentes de cada tópico en el Anexo IV. En esa sección además, se presentan nombres tentativos a cada uno de los tópicos, que se obtuvo por medio de revisión de textos con esos tópicos. A continuación se presenta el número de interacciones por tópico.

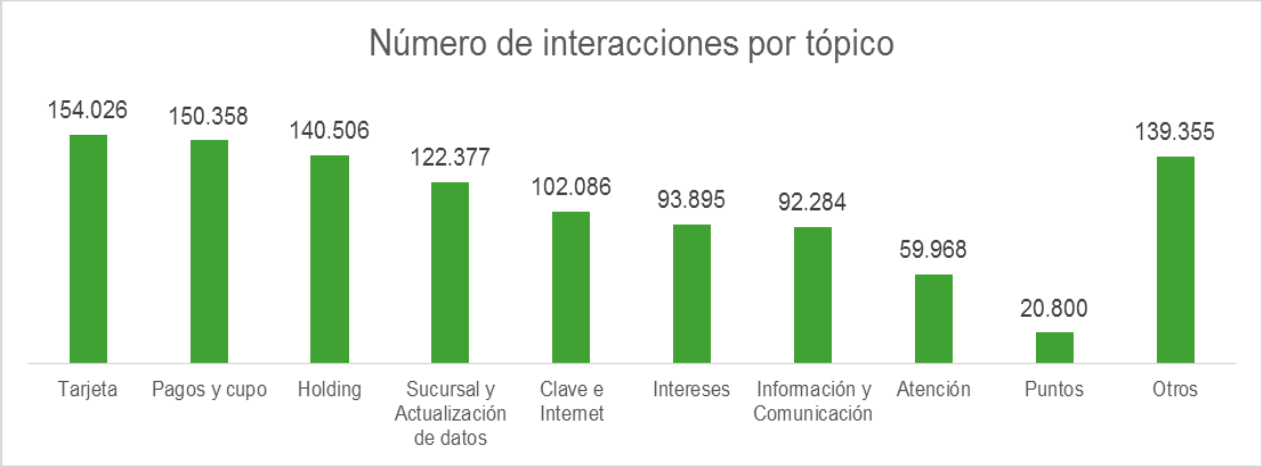


Grafico 13: Numero de interacciones por tópico
Fuente: Elaboración propia con datos de la compañía

Se puede ver en el gráfico, las principales interacciones se deben a interacciones por la tarjeta de la compañía, cupos y pagos, y comentarios del holding en general. Por el otro lado hay pocos comentarios de los puntos, la atención por el servicio y de información de la empresa. Es importante destacar que, si bien se podría decir que pagos y cupos son elementos de la tarjeta, es tenerlos separados como dos tópicos permite ver los diferentes elementos de cada uno.

7.2.4 Análisis de sentimientos

Se implementó el algoritmo presentado en la sección de marco conceptual. Este se utilizó con los diccionarios de la memoria de Víctor Cortes [13] disponibles en español. Los resultados obtenidos fueron los presentados en el siguiente gráfico.

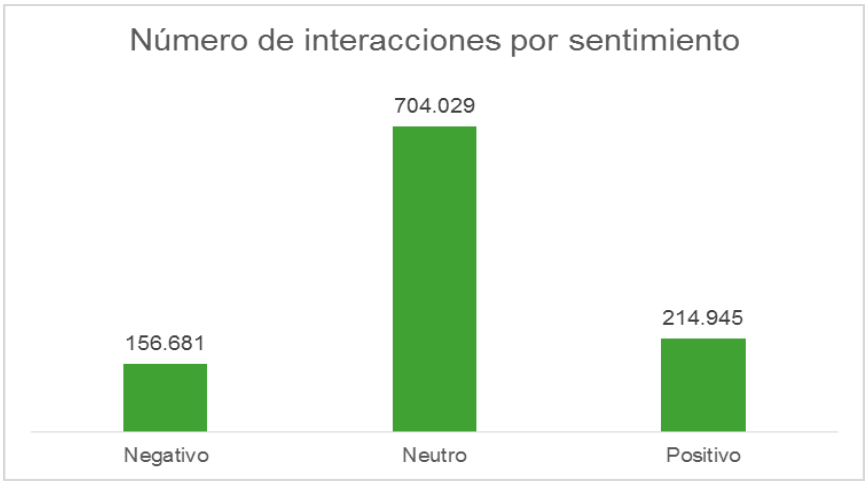


Grafico 14: Numero de interacciones por sentimiento
Fuente: Elaboración propia con datos de la compañía

De este grafico se puede ver que la mayoría de los comentarios para la empresa son neutros. Esto hace sentido considerando que es el trato con una empresa, y que cierto grado de formalidad es necesario. Sin embargo, existe un porcentaje importante de comentarios que cuentan con aspectos positivos. Esto puede ser por la gran cantidad de datos de nps, que es un indicador importante dentro de la empresa y que suele estar positivo. En último lugar se tienen los comentarios negativos.

De evaluarse de manera individual los sentimientos, la menor cantidad de comentarios con sentimientos positivos y negativos en comparación a los comentarios con un sentimiento neutro podría presentar un problema por la poca varianza de los valores. Sin embargo, dado que las variables de tópicos y de sentimientos se tomarían en conjunto para cada comentario, esto no presentaría un problema.

Por último, se realiza un cruce entre los datos por tópico y sentimiento respectivamente.

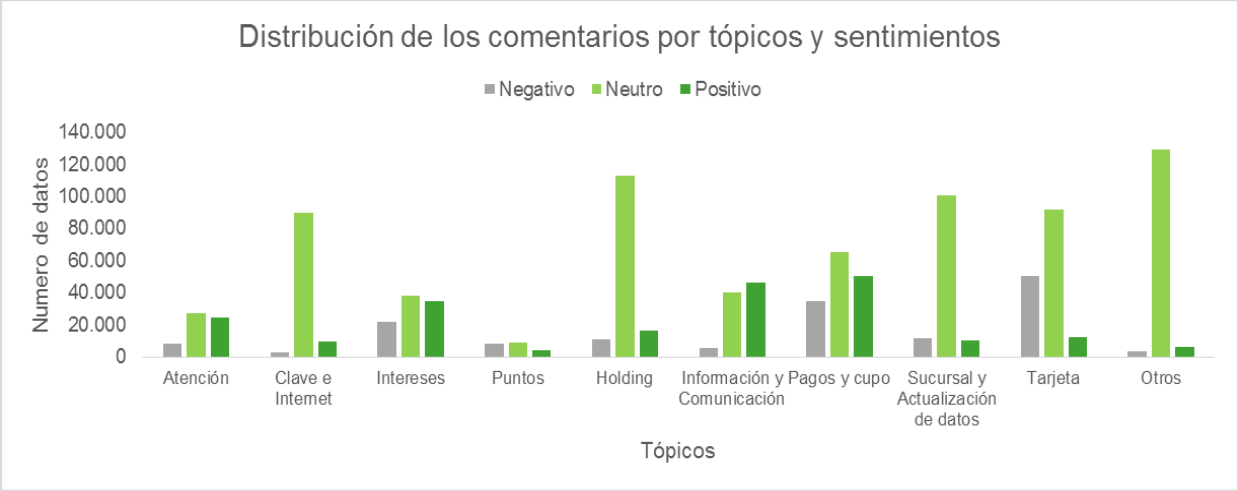


Grafico 15: Número de interacciones separado por tópicos y sentimientos
Fuente: Elaboración propia con datos de la compañía

Tópico	%Negativo	%Neutro	%Positivos	%Total	# Total
Tarjeta	33%	60%	8%	100%	154.026
Pagos y cupo	23%	44%	33%	100%	150.358
Holding	8%	80%	12%	100%	140.506
Otros	3%	93%	5%	100%	139.355
Sucursal y Actualización de datos	9%	82%	9%	100%	122.377
Clave e Internet	3%	88%	9%	100%	102.086
Intereses	23%	40%	37%	100%	93.895
Información y Comunicación	6%	44%	50%	100%	92.284
Atención	14%	46%	41%	100%	59.968
Puntos	38%	41%	20%	100%	20.800

Tabla 8: Concentración de comentarios por sentimientos y tópicos
Fuente: Elaboración propia con datos de la compañía

El grafico y la tabla anterior se pueden desprender como son vistos determinados elementos de la compañía por parte de los clientes. Recordando que los tópicos con más comentarios eran los de tarjeta, cupo y pago, y por ultimo holding, Es interesante notar que un porcentaje importante de los comentarios asociados a la tarjeta son negativos. Esto presenta una gran oportunidad de gestión dado que logra señalar en particular los ámbitos donde se pueden tener actividades para mejorar la experiencia del cliente, por medio de las palabras en cada uno de los tópicos (Estas últimas se pueden ver en el Anexo IV). Al mismo tiempo, el área de pagos y cupos por ejemplo presenta una importante cantidad de datos positivos y negativos. Por medio de estos tópicos y sentimientos se podría detectar las fortalezas que sienten los clientes tiene la tarjeta, y explotarla dentro de campañas de marketing de mejor manera. Es interesante por ultimo notar que los tópicos con la mayor cantidad de datos parecieran tener una connotación mes negativa que tópicos que se encuentran con una menor cantidad de datos, como es el caso de intereses, información/comunicación y atención.

7.3 Homologación de clientes

7.3.1 Modelo de propensión a tópico

Este modelo, como se explicó anteriormente, fue construido en base a un logit multivariado. La variable objetivo de este logit fue una marca conjunta de tópicos y sentimientos asociadas a un determinado comentario. Para lograr calibrar el modelo se tomaron las interacciones de la base y se obtuvieron los siguientes resultados.

Información de ajuste de los modelos				
Modelo	Criterios de ajuste de modelo	Pruebas de la razón de verosimilitud		
	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	1,449,010,263			
Final	1,417,184,037	31,826,226	696	0

Tabla 9: Significancia general del modelo de propensión a comentario
Fuente: Elaboración propia con datos de la compañía

De este grafico se desprende que el modelo en si es algo valido para predecir el comportamiento, dado que tiene una significancia igual a cero.

7.3.2 Resultados del emparejamiento

Una vez realizado el modelo de propensión comentario, se utiliza un algoritmo de KNN para lograr identificar a los gemelos presentes dentro de la base de datos que no tienen

comentario. Esto se realiza con el objetivo de lograr emparejar a los clientes y determinar el efecto que tienen en el comportamiento. Para esto, se seleccionan a los 3 vecinos más cercanos.

Una vez que se tienen seleccionados los 3 vecinos más cercanos, se calcula el promedio de la variación en el gasto con respecto a un año anterior y después de la realización del comentario. Lo mismo se realiza con el cliente que si realiza el comentario. Esto se hace con el objetivo de encontrar cuanto es la variación neta que hace el comentario. Es importante destacar que no todos estos clientes tienen gasto un año antes o un año después del momento al que se le compara, por lo que se seleccionan los 3 vecinos más cercanos y se obtiene un promedio de los gastos. La variación final por comentario se presenta a continuación.

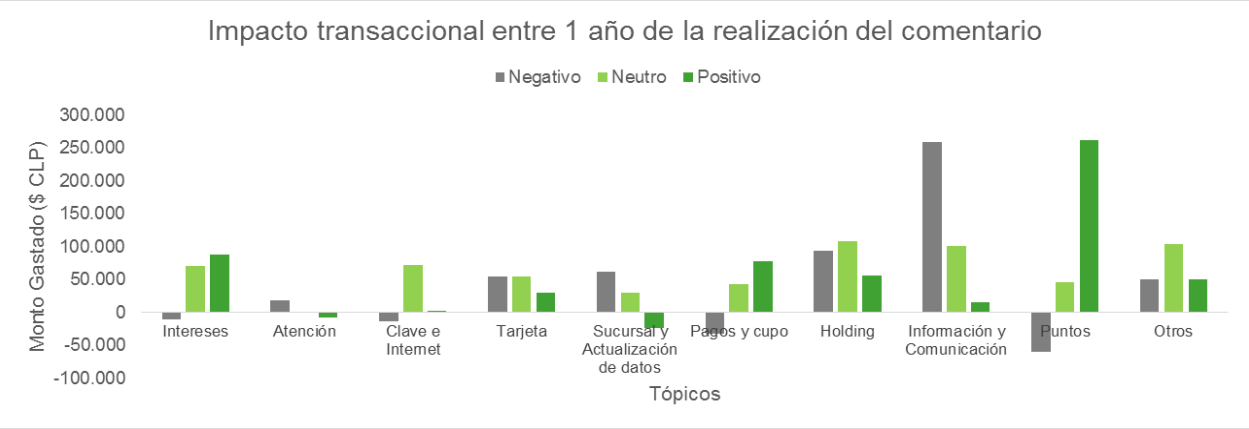


Grafico 16: Significancia general del modelo de propensión a comentario
Fuente: Elaboración propia con datos de la compañía

De este grafico es posible apreciar 2 tipos de impactos entre los clientes que realizan un determinado comentario. Por una parte, existes ciertos tópicos (como los de otros, tarjetas y holding) que tienen un efecto positivo en el monto gastado por los clientes un año después sin importar el sentimiento que estos tengan. Por otra parte, existen otros tópicos (como los de puntos, pagos y cupos e intereses) en que el sentimiento tiene un efecto multiplicador dentro del gasto que realizan estos clientes un año después, gastando más si es positivo y menos si es negativo .

Dentro de los 2 tipos de impactos posibles, se pueden ver diferentes grados de impacto para diferentes tópicos.

Uno de los puntos más interesantes obtenidos de este modelo es el efecto que tiene el tópico de puntos dentro del comportamiento de los clientes. Al pertenecer a los tipos de comentarios que multiplican el gasto dependiendo del sentimiento, este tópico logra aumentar unas 2,4 veces el gasto en caso de ser positivo con respecto al tópico de pagos/cupo, que tienen el mismo impacto y tiene uno de los efectos más grandes en el gasto promedio de sus clientes a futuro. Esto deja a que las personas que emiten comentarios de puntos gaste en promedio

$262.263 pesos chilenos más un año después de la realización del comentario. Al mismo tiempo, el tópico de puntos es un tópico que disminuye en un 84% más el impacto del tópico de pagos/cupos para los comentarios negativos, con lo que un cliente que emite un comentario negativo gastaría $60.097 pesos chilenos.

Esto es interesan de dado el gran impacto que tienen, y a la pequeña cantidad de comentario que existen de este tópico. Al ser una cantidad de comentarios pequeño, una estrategia personalizada podría ser útil para los clientes que emiten este tipo de comentarios. Esto se contrapone a lo que habría que hacer con los clientes que emiten comentarios de Pagos/Cupos, dado que estos son una cantidad considerablemente mayor a la de los comentarios con puntos (129.558 comentarios más que el tópico de puntos). Posiblemente para estos comentarios puede ser una buena estrategia la iteración de la clasificación de tópicos para esta sub-muestra de datos. Por medio de esto, se podría saber cuáles son los elementos comunes dentro de estos comentarios para planificar estrategias corporativas que logren potencias el efecto positivo, y disminuir el negativo.

Con respecto a los comentarios que generan un aumento luego del comentario sin importar el sentimiento, una estrategia corporativa sería difícil de implementar con la información obtenida. Al igual que con el tópico de pagos/cupo, puede ser una buena estrategia la iteración de la clasificación para descubrir si existe un elemento común que permita construir una estrategia robusta, y que logre verdaderamente diferenciar cuales son los tópicos que agradan y no a los clientes dentro de los comentarios que tienen este tipo de impacto. Esto es posible en particular con los comentarios de los tópicos de holding y otros, dado que existen grandes cantidades de estos comentarios.

7.4 Modelo de fuga

7.4.1 Unificación de la base de datos

Una vez que se tuvo los sentimientos y tópicos dentro de la tabla de interacciones, se puede unificar la data de los clientes con la data de los comentarios. Dado que el identificador dentro de la tabla de interacciones es el identificador de la interacción, la tabla final también queda identificada por estos datos, donde aquellos clientes sin comentarios tendrán celdas vacías dentro de estos valores.

7.4.2 Eliminación de Outliers

Se consideran como outliers aquel dato que sale del comportamiento normal de los otros registros dentro de una base de datos. Dentro del contexto de una tarjeta de crédito, se consideró como variables que podrían detectar un valor anómalo como la frecuencia de compra que tiene algún cliente o los montos que gasta con su tarjeta.

Para poder eliminar valores simultáneamente teniendo en consideración estas dos variables se decidió utilizar la distancia de Mahalanobis por sobre el conjunto de datos. Aquellos

datos que se encontrasen por fuera de cierto radio serian entonces los registros eliminados. Es importante además recalcar que se utiliza un logaritmo para encontrar estos datos extremos dado que normaliza los valores, dadas las escalas que se tienen

En base a esto, se logró eliminar de la base aproximadamente un 3% de los datos, donde una representación gráfica de la eliminación se puede ver a continuación:

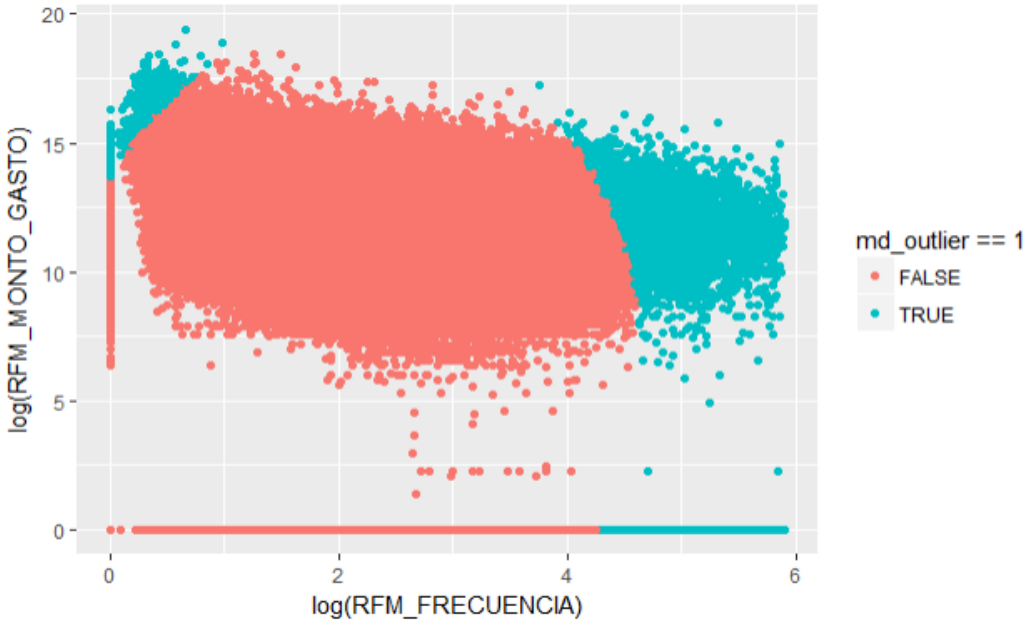


Grafico 17: Eliminación de outliers por medio de la distancia de Mahalanobis
Fuente: Elaboración propia con datos de la compañía

7.4.3 Balanceo de la base de datos

Para lograr tener el mayor efecto predictivo posible, fue necesario balancear las bases de datos disponibles. Esto se debe a que existía una diferencia considerable entre los datos con comentario escrito (249.200) y aquellos sin comentario escrito (1.047.303).

Para esto se utilizó la técnica de over-sampling. Esto se decidió dado que si bien el balanceo por medio de técnicas de la generación sintética de registros (SMOTE) presenta un mejor desempeño, presentaba un problema a la hora de homologar clientes. Dada la naturaleza sintética de los datos generados por SMOTE, pudiera tenerse una diferencia en el monto real gastado, entre las personas con y sin comentario, al identificarse erróneamente al cliente, y por ende, hacer un mal cálculo del vecino más cercano. Finalmente, después del balance se quedó con 2.094.606 datos en la base final.

7.4.4 Selección de variables

Dado que se la base de datos contaba con 95 variables para los clientes, se desarrolló una selección de variables por medio de un análisis de correlaciones y un análisis de como las variables lograban explicar la fuga de los clientes. En ambos casos se siguió la metodología planteada en artículo [16].

1.- Análisis de correlación:

El análisis de correlaciones planteado en el paper antes mencionado consiste en lograr calcular la suma de valores absolutos de cada uno de las variables a considerar. Luego de esto se define un promedio entre todos estos valores. Luego se define un límite k donde

$$k = t * \text{correlacion promedio} \quad \text{donde t} \in \{0,1,..\}$$

Ecuación 11: Ecuación para el cálculo de variables por correlación

Una vez que se define el k, se eliminan todas las variables que tengan una correlación mayor a k.

2.- Information Gain

Information gain es una medida de entropía, que logra calcular cuanta información entrega una determinada variable a la variable dependiente. Fue presentada en con más detalle en el marco teórico. Por medio de estas 2 técnicas, se logró quedar con las variables.

Variable	Considerar variable de 3 meses anteriores	Nombre de la variable
Tipo de tarjeta del cliente		ID_MULTIPROD
Sexo del cliente		ID_SEXO
PGI en comunicaciones	Si	SOW_COMUNICACIONES
PGI en recaudaciones	Si	SOW_RECAUDACION
PGI en seguros	Si	SOW_SEGUROS
PGI en tiendas por departamento	Si	SOW_TIENDAS_DPTO
PGI en transporte	Si	SOW_TRANSPORTE
Numero de rubros en que compra el cliente fuera del holding	Si	NUM_RUBROS_OT
Suma del monto gastado fuera del holding	Si	SUM_MONTO_OT
Variación porcentual del monto gastado con la tarjeta en los últimos 3 meses		VAR_POR_MONTO_GASTO
Variación porcentual en el PGI de la tarjeta en los últimos 3 meses		VAR_SOW
Variación porcentual en el puntaje de riesgo del cliente en los últimos 3 meses		VAR_SCORE
Variación porcentual del monto gastado fuera del holding con la tarjeta en los últimos 3 meses		VAR_MONTO_OT
Marca si compra en tiendas por departamento del holding		RETAIL
Deuda que tiene el cliente		DEUDA
Cantidad de puntos acumulados por el cliente		F_PUNT_ACUM
Porcentaje de saturación del cupo del cliente		SATURACION_CUPO
Marca si el cliente es de uso frecuente de la tarjeta	Si	ENGANCHADO_TARJETA
Dinero disponibles en superavances		DISPONIBLE_AVANCE
Dinero disponible para el cliente en efectivo		DISPONIBLE_SAV
Marca conjunta del tópico y sentimiento de un determinado comentario obtenido con los algoritmos de minería de texto		GRUPO_COMENTARIO

Ilustración 11: Selección de variables por medio de information gain y análisis de correlación
Fuente: Elaboración propia con datos de la compañía

Las otras variables consideradas cumplen con las siguientes 2 condiciones:

- Tiene una correlación total con las otras variables de la base inferior a dos veces el promedio de la base con todas las variables
- Tiene un valor de information gain mayor a cero

7.4.5 Modelo de fuga tradicional

Variables más importantes

Las variables que mejor logran explicar la fuga de los clientes se explica por medio del siguiente gráfico.

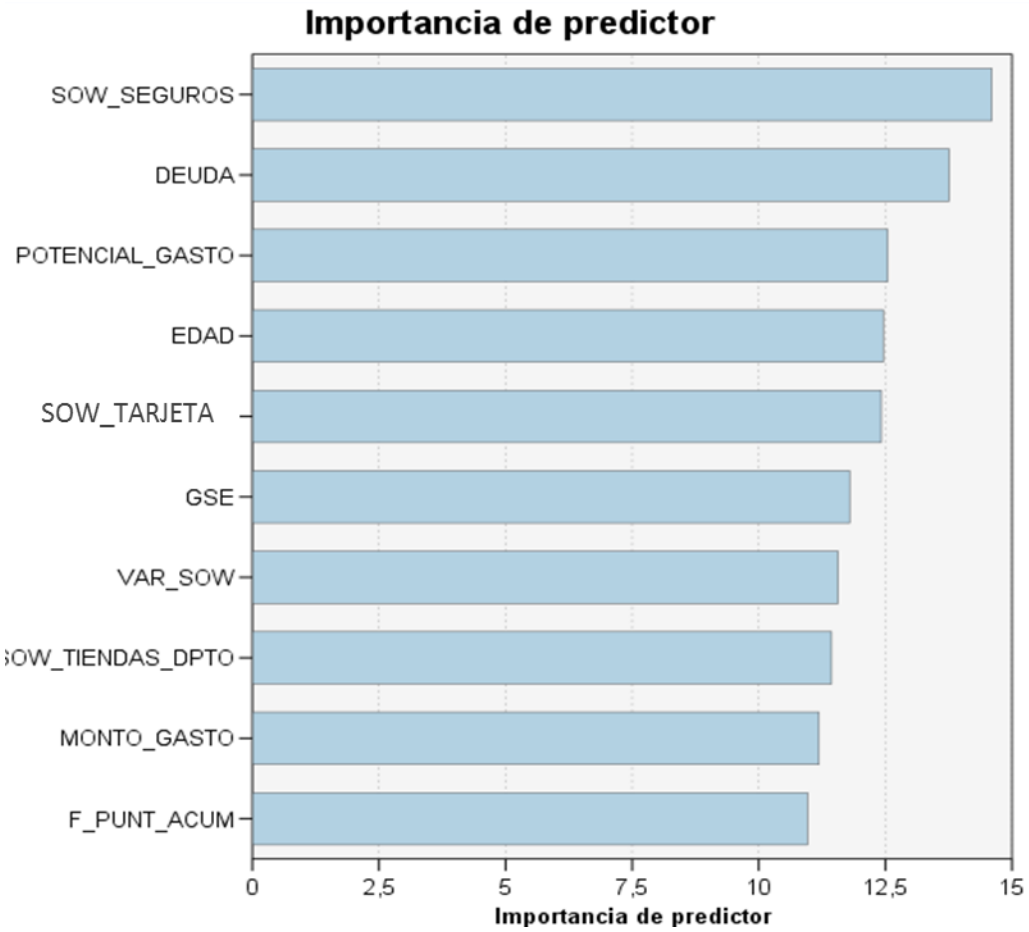


Ilustración 12: Selección de variables más importantes con el random forest en modelo tradicional de fuga
Fuente: Elaboración propia con datos de la compañía

Este grafico señala que para la predicción de fuga los aspectos importantes actúan de manera relativamente homogénea para un cliente. Dentro de las variables más importantes se puede ver que son principalmente variables de comportamiento del cliente, en el sentido de que las única variable de tenencia de productos va asociada a los puntos acumulados, y la únicas variables demográficas importantes son las de edad y ges del cliente.

Reglas más comunes

Las reglas más comunes por otra parte son instrucciones que se desprenden de los árboles en base a determinadas condiciones que cumplen las variables y que indican una alta probabilidad del cliente a fugarse o a no fugarse. Este resultado para el random forest se puede ver en el Anexo V.

Estudiando las reglas más comunes para los clientes fugados, se puede ver lo siguiente:

- El grupo más propenso a la fuga es un grupo de clientes que ya dejo de realizar gastos con la tarjeta. Esto se ve reflejado en el hecho de que son clientes que disminuye su gasto con respecto a los últimos 3 meses casi en su totalidad, y que además no compra fuera del holding en los últimos 3 meses en los rubros fuera del holding en los últimos 3 meses. Dado que 20% de las personas que cumplen esta condición se fuga (Precisión de la regla en la tabla), es un aspecto importante a considerar.
- Por otra parte, las dos reglas siguientes son más complejas de analizar. Esto se debe a que consideran la variación del puntaje de riesgo de los clientes. No se puede determinar si el aumento en el riesgo del cliente causa la fuga o si es porque como el cliente cambia su comportamiento al fugarse esto se ve reflejado en la variable del puntaje de riesgo. Sin embargo, para ambos grupos son clientes que tienen poco gasto. Para el segundo grupo los puntos acumulados no superan los 1613 puntos, y para el tercero no compran en empresas fuera del holding, al igual que el primer grupo descrito. Para el segundo grupo la precisión de la regla es de un 19% y la tercera tiene una precisión del 18,2%.

El presente bosque, al igual que todos los bosques que siguen, no presenta reglas comunes para los no fugados, por lo que no se incluye esta información.

Desempeño del modelo

Por último, se presenta el desempeño que tiene este modelo con los datos analizados.

	Entrenamiento	Comprobación
Accuracy	0,78	0,78
Precisión	0,85	0,82
AUC	0,87	0,85

Tabla 10: Tabla de desempeño del modelo de fuga tradicional
Fuente: Elaboración propia con datos de la compañía

7.4.6 Modelo de fuga con variables de minería de texto

Para poder abordar este problema de mejor manera se crearon 2 modelos con minería de texto. El primero de ellos consideraba solo a clientes que tenían interacciones escritas. El segundo se consideró con el total de la base para lograr ver cuál sería el efecto de las variables de minería de texto dentro del modelo. Esto se realizó para lograr identificar si los clientes sin comentarios generaban algún ruido con respecto a los otros datos.

Modelo con variables de minería de texto exclusivamente

Con respecto a las variables del mayor impacto entregado por el random forest, el resultado se presenta a continuación.

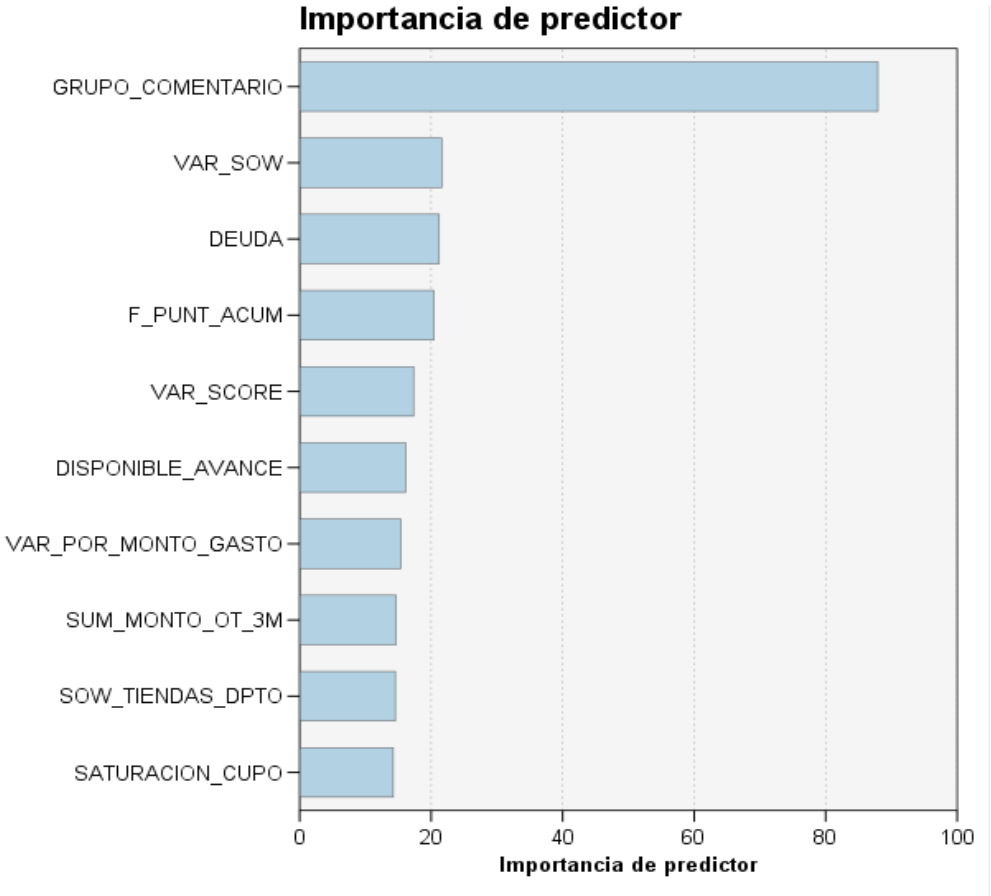


Ilustración 13: Variables más importantes con el random forest en modelo de fuga con variables de minería de texto exclusivamente considerando solamente a los clientes con comentarios
Fuente: Elaboración propia con datos de la compañía

Dentro de este grupo de datos, se puede ver que el impacto de la variable grupo comentario es considerablemente mayor al de las otras variables estudiadas dentro de este conjunto de datos. Esta variable es la generada en base a una marca conjunta de Tópicos y Sentimientos obtenidos de los modelos de minería de texto. Entre las otras variables importantes que resultan de random forest, logra compartir cuatro variables con el modelo anterior, que serían las de deuda, puntos acumulados, PGI de tiendas por departamento y la variación del PGI.

Con respecto a las reglas más relevantes, estas se pueden ver en el Anexo VI. Se puede ver en primera instancia que la precisión de la regla es inferiores a las del modelo anterior (las precisiones de las reglas son de 18,7% 15,6% y 17,6% respectivamente para los grupos 1,2 y 3). Además de esto, se tiene que la saturación de cupo, el enganche en la tarjeta y el número de rubros en el que compra fuera del holding afecta a la fuga de estos clientes. Si el cliente no está enganchado con la tarjeta, no compra en rubros en empresas del holding y una saturación de cupo baja, es muy probable que este cliente se fuga.

El primer grupo por ejemplo no presenta gran diferencia entre su disponible y su cupo (saturación baja), no es un cliente frecuente de la tarjeta. Sin embargo, es interesante considerar que son clientes con una PGI de seguros menor a 77% de sus ingresos, y que tiene una deuda inferior a $77.144 pesos chilenos. Esto perece ser interesante dado que muestran a un cliente que gastaba un porcentaje de sus ingresos con la tarjeta, a diferencia de los clientes que mostraban una señal previa tres meses antes de fugarse.

. El segundo grupo también presenta la saturación baja, tienen menos de 2495 puntos y no compro en tiendas fuera del holding en los últimos 3 meses. Al tener pocos puntos, se puede ver que es un cliente que usualmente usa poco la tarjeta.

El tercer grupo no es un cliente que use frecuentemente la tarjeta, pero sorpresivamente si puede gastar dinero en tiendas fuera del holding, lo que se ve reflejado en el monto y cantidad de tiendas en las que compran fuera del holding. Dadas estas características, podría ser un cliente que encontró una mejor alternativa dentro de otra empresa del mismo rubro. Esto también es apoyado por el hecho de que tres meses antes de la fecha estudiada no compraba fuera del holding y si destinaba una parte de su dinero a seguros.

Por último, las métricas de desempeño para este modelo se muestran a continuación.

	Entrenamiento	Comprobación
Accuracy	0,79	0,78
Precisión	0,88	0,81
AUC	0,90	0,85

Tabla 11: Métricas de desempeño del modelo de fuga con variables de minería de texto exclusivamente considerando solamente a los clientes con comentarios
Fuente: Elaboración propia con datos de la compañía

Se puede ver que mantiene parámetros relativamente similares con respecto al grupo de entrenamiento y de comprobación.

Modelo con variables de minería de texto considerando al total de los clientes

La importancia relativa de los indicadores entregados por este modelo se presenta a continuación.

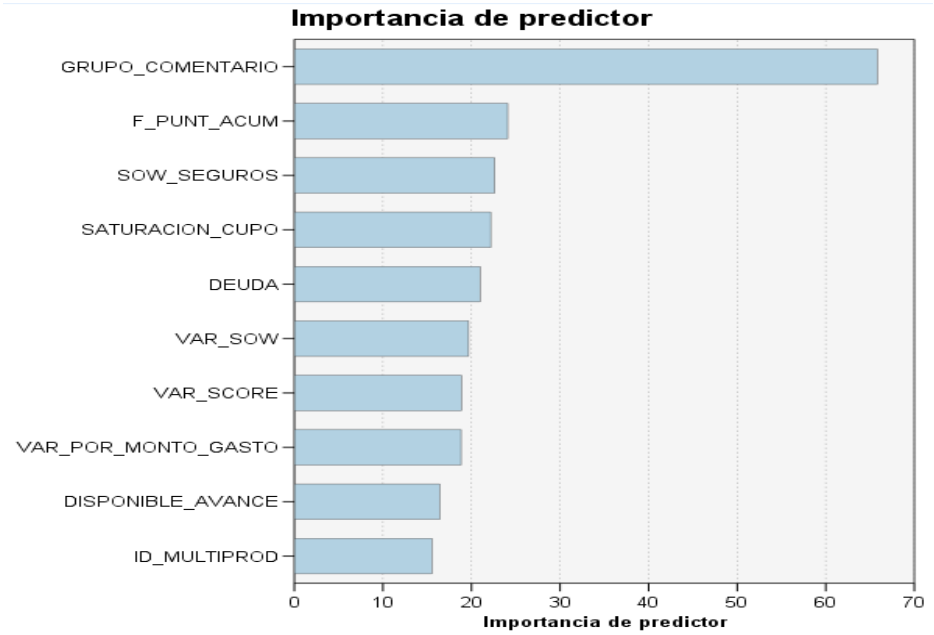


Ilustración 14: Métricas de desempeño del modelo de fuga con variables de minería de texto exclusivamente considerando solamente a los clientes con comentarios
Fuente: Elaboración propia con datos de la compañía

Se sigue observando que la variable del grupo del comentario es un predictor importante dentro de este modelo. Aun así, cabe destacar que si bien es el mejor predictor dentro de las variables consideradas para este modelo, tiene una importancia mucho menor a la importancia presentada por la misma variable para el modelo anterior. Esto se puede ver en la escala de la importancia del predictor para ambos casos.

Con respecto a las reglas más comunes presentadas para este modelo, estas se presentan en el Anexo VII. El primero de estos grupos llama poderosamente la atención. Estudiando primero las características propias de este cliente, vemos que este era un cliente que no utilizaba mucho la tarjeta en un pasado. Esto se puede ver en el hecho de que el cliente tiene pocos puntos acumulados. También se puede ver que es un cliente que no gasta mucho dinero en empresas fuera del holding, por el número de rubros en los que compra fuera del negocio. Sin embargo, es un cliente que empieza a gastar más dinero en la empresa, y emite comentarios del servicio prestado por la empresa. Este sub-conjunto de comentarios es muy interesante dado que está asociado a un cliente que re-activa su uso de la tarjeta. El profundizar en los aspectos de estos comentarios podría ser una buena alternativa para saber cuáles son los aspectos que llaman más la atención a clientes entrantes.

El segundo grupo no presenta una diferencia importante entre su cupo y su dinero disponible. Además de esto no suele gastar en recaudación. Se puede asumir que este no es un cliente que utilice mucho la tarjeta.

Por último, el tercer grupo suele tener también una baja saturación de cupo. Sin embargo, es un cliente que puede gastar mucho en tiendas por departamento, por su PGI y puede gastar mucho dinero en tiendas fuera del holding. El retener a un cliente que cumpla con estas características es muy valioso para la compañía, dado que son clientes que gastan más.

Por último, con respecto al desempeño de este modelo, tenemos que su desempeño con respecto a las métricas estudiadas es el siguiente

	Entrenamiento	Comprobación
Accuracy	0,76	0,76
Precisión	0,85	0,98
AUC	0,86	0,84

Tabla 12: Métricas de desempeño del modelo de fuga con variables de minería de texto considerando a todos los clientes
Fuente: Elaboración propia con datos de la compañía

7.4.7 Modelo de fuga con clientes homologados

Para lograr extrapolar la información de los clientes que tenían comentarios y transferírsela a los clientes sin comentarios, se optó por la utilización del resultado obtenido en base a logit multivariado utilizado en la sección de homologación de clientes. Si bien este modelo no es perfecto por la necesidad de introducir la variable resultante dentro del modelo de fuga como se explicó con anterioridad, se pretende ver si esto sería suficiente para lograr un modelo de fuga mejor que los presentados anteriormente.

Es importante considerar que las variables utilizadas dentro del logit multivariado no podrían ser utilizadas dentro del modelo de fuga, dada la alta correlación de variables que se generaría con la nueva variable generada.

Las principales variables para el modelo de random forest son las siguientes.

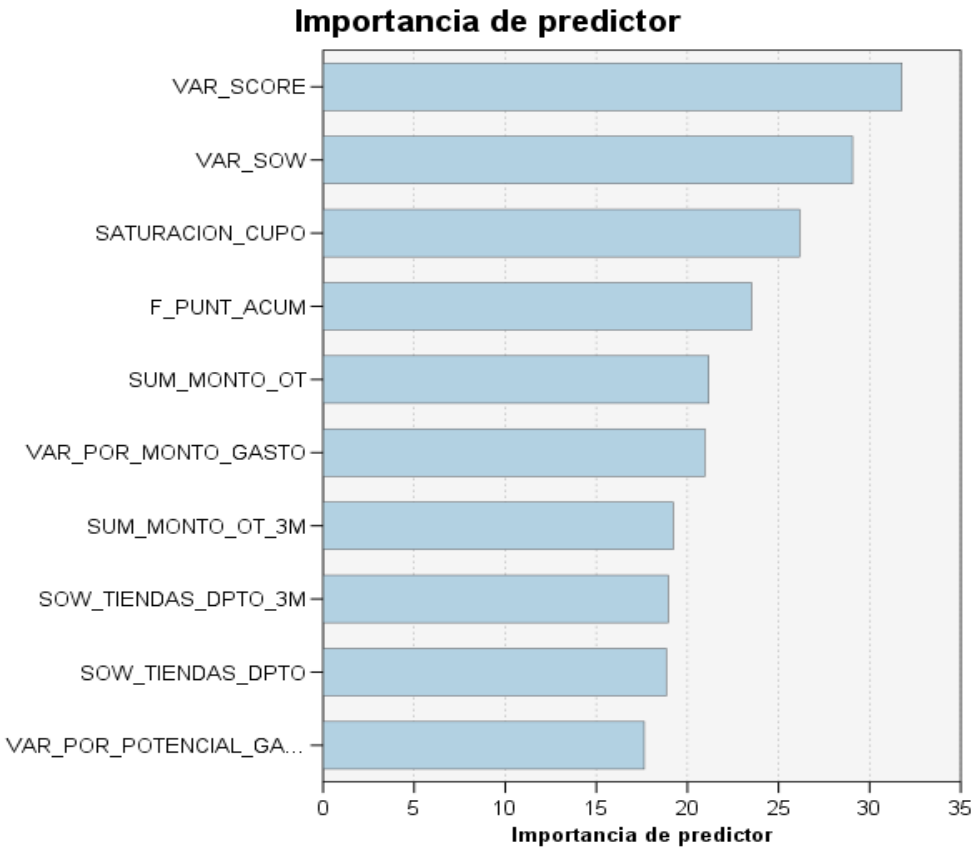


Ilustración 15: Métricas de desempeño del modelo de fuga con variables de minería de texto generadas en base al logit multivariado de la homologación de clientes
Fuente: Elaboración propia con datos de la compañía

Este resultado, a diferencia de otros presentados anteriormente, da mucha importancia a las variables de variación porcentual del comportamiento de los clientes. Esto se debe a que debe remplazar el impacto de algunas de estas por las variaciones porcentuales del mismo indicados. Es importante destacar también que la nueva variable generada no está considerada entre las 10 variables de más impacto del modelo

Con respecto a las reglas más relevantes, estas se presentan en el Anexo VIII. La regla más interesante de estudiar es la del segundo grupo. Al estudiar su comportamiento 3 meses antes de realizar el análisis, este cliente solía dedicar como máximo un 40% de su dinero a tiendas por departamento, un 25% como máximo de su dinero a seguros y gastaba un máximo de $190.340 pesos en tiendas fuera del holding. Esto nos señala que es un cliente que hacía uso de la tarjeta en los pasados 3 meses. Dada estas característica, y que en el momento estudiado no realiza compras fuera del holding, podría ser una buena alternativa la de presentarle ofertas para no perder su interés en la tarjeta.

Por último el desempeño del modelo se presenta a continuación.

	Entrenamiento	Comprobación
Accuracy	0,76	0,76
Precisión	0,87	0,82
AUC	0,87	0,84

Tabla 13: Métricas de desempeño del modelo de fuga con variables de minería de texto generadas en base a logit multivariado del modelo de homologación de clientes
Fuente: Elaboración propia con datos de la compañía

7.4.8 Selección de modelo

Para lograr seleccionar el mejor modelo disponible entre los datos, se pasa a ver el desempeño de cada uno de los modelos con su respectivo grupo de comprobación.

	Fuga Tradicional	Fuga tradicional con variables de minería de texto (solo con clientes con comentarios)	Fuga tradicional con variables de minería de texto (todos los clientes sin homologar)	Fuga tradicional con variables de minería de texto (todos los clientes con homologación)
Accuracy	0,775	0,779	0,761	0,757
Precisión	0,820	0,809	0,977	0,820
AUC	0,851	0,854	0,837	0,842

Tabla 14: Métricas de desempeño en los grupos de comprobación para cada uno de los modelos
Fuente: Elaboración propia con datos de la compañía

Para poder seleccionar el mejor modelo es necesario primero comprender cuales son las 3 métricas que se están estudiando. Accuracy es una métrica que calcula del total de datos que porcentaje fue clasificado de manera correcta. Por otra parte, la precisión presenta el porcentaje de los datos predichos como fugados, que porcentaje efectivamente debe ser clasificado como fugado. Por último, El AUC es el área bajo las curvas ROC, que están construidas para poder determinar cuanto mejor es un determinado modelo al lanzamiento aleatorio de una moneda.

Estudiando los estadísticos de manera general, se puede ver que el modelo de fuga con variables de minería de texto solo para clientes con comentario presenta el mejor desempeño para las métricas de accuracy y AUC. La diferencia de estas métricas con el modelo de fuga tradicional es bastante baja (menos de 0,004 en ambas métricas). Pero el modelo de minería de texto tendría la ventaja en este caso de tener los tópicos de mayor interés para los clientes y se podría gestionar en base a sus necesidades. Al estudiar la precisión se puede ver que el mejor modelo es el con toda la cartera de clientes y considerando las variables de minería de texto pero sin considerar la homologación de comentarios. Con respecto a esta misma métrica se puede ver que el modelo que le sigue es el de fuga tradicional, y posteriormente el modelo de fuga con variable de minería de texto con la cartera de clientes con comentarios.

En base a esto se debería seleccionar el modelo que considera las variables de minería de texto solo para los clientes con comentarios. Esto se debe a que con respecto al modelo de fuga tradicional, este último tiene un peor desempeño en las métricas de accuracy y AUC por 0,4% y

0,3%. Sin embargo, este modelo solo puede ser corrido para clientes que tienen un comentario asociado. Esto, al no poder ser generalizable, no puede ser utilizado como modelo final dentro de la compañía.

Analizando el modelo de fuga tradicional se puede ver que tiene un mejor desempeño que el modelo de fuga con variables de minería de texto de toda la cartera pero sin homologar. Con respecto a la accuracy y de AUC tiene un desempeño superior por 1,4%. Es interesante comparar esto modelos con respecto a la precisión, dado que el alto grado de precisión dentro del modelo sin homologar y peor desempeño general apunta a que el error se centra en clientes que se dicen no fugados y después si se fugan. Dentro del contexto de la fuga esto es un tema de mayor riesgo, por lo que dado esto y es mejor desempeño en las otras métricas se decide seleccionar el modelo de fuga tradicional.

Al comparar los cuatro modelos presentados anteriormente existen dos conclusiones importantes por destacar. Por una parte, se puede ver que dado el mejor desempeño del modelo con variables de minería de texto y la cartera de clientes con comentarios, las variables de minería de texto si aportan valor para la predicción de la fuga. Sin embargo, esto solo sirve dentro del contexto en que todos los clientes tengan un comentario. Es por este motivo que el modelo de fuga tradicional tiene mejor rendimiento que el modelo con las variables de minería de texto sin homologar y con toda la cartera. La otra conclusión importante es que la homologación de comentarios por medio del logit multivariado, también utilizado en el propensity score matching no es la mejor manera de extrapolar los intereses de determinados clientes, motivo por el cual fue el algoritmo de peor rendimiento general.

7.4.9 Análisis de Sensibilidad

Para la realización del análisis de sensibilidad se tomaron los datos de la predicción del modelo original, y se determinó cual era la probabilidad a la fuga que tenía cada uno de los clientes. En base a esto, se fue modificando el punto de corte del algoritmo, para poder así ver que ocurría con las diferentes métricas a medida que se movía este punto de corte.

Para poder hacer esta revisión de manera completa, se estudió al mismo tiempo los errores de tipo I y II que pudiesen ocurrir. Recordando que el error de tipo I se define como aquel en que se predice que un resultado sea negativo (en este caso que el cliente no se fuga) y se tiene que el resultado es positivo (el cliente se fuga). Al mismo tiempo, el error de tipo II consiste en que se predice que el resultado va a ser positivo, y el resultado es negativo. Los resultados de este análisis se presentan en el siguiente gráfico.

Para poder realizar el análisis de sensibilidad, es necesario definir un punto de corte para la asignación de fuga. El punto de corte se define como el punto donde a partir del cual se define la asignación de fuga o no fuga en base a la propensión. Por ejemplo, el punto de corte es mayor a la propensión se asigna como no fugado. Por otra parte, si se cumple la condición contraria se define como fugado. El resultado del análisis para diferentes métricas y puntos de cortes se presenta a continuación.

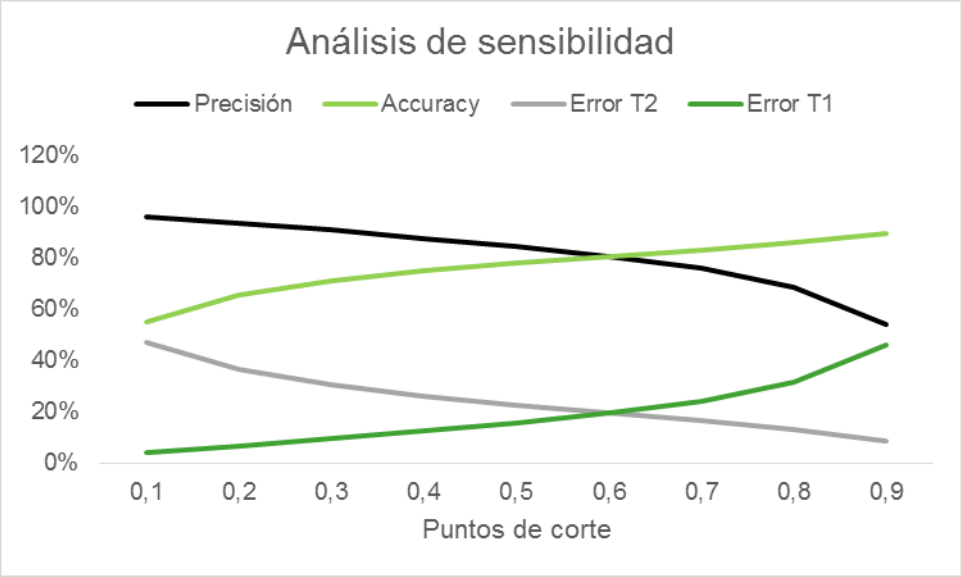


Grafico 18: Métricas de desempeño con respecto al punto de corte de la asignación del modelo de fuga tradicional
Fuente: Elaboración propia con datos de la compañía

Dados los resultados del análisis de sensibilidad se puede ver, en primera instancia que el error de tipo I y II se minimizan en diferentes extremos de los puntos de corte. El error total por otra parte, se minimiza en 0,6. Por otra parte, la precisión y accuracy presentan un problema similar. Ambos se maximizan en diferentes extremos de los puntos de corte de asignación. Por lo tanto ambos se maximizan nuevamente en 0,6. Es por esto que se decide cambiar la asignación de 0,6 como punto de corte.

7.5 Trabajos futuros

Uno de los puntos donde se podría profundizar los temas tocados es en el cómo afectan las palabras y sus probabilidades dentro del comportamiento transaccional de los clientes de cada tópico. El poder determinar cuáles palabras tienen un impacto positivo y negativo podrían dar luces de manera más clara de cómo lograr armar una estrategia para empresas.

Como se pudo apreciar en los resultados obtenidos de la memoria, la técnica logit multivariado no presentaba una buena alternativa para señalar cuales son los tópicos que interesan a determinados clientes en el modelo homologado. Pudiese ser interesante en otro tipo de metodología o experimento que lograse determinar cómo se podría predecir estas variables para clientes que no las tienen.

7.6 Líneas de acción a proponer

Dados los resultados obtenidos a partir de la memoria, se tienen las siguientes recomendaciones para la empresa.

Por una parte, dado lo descubierto en el impacto transaccional de los dos tipos de impacto que tienen los comentarios escritos se recomienda que:

- Para los tópicos que tienen una pequeña cantidad de datos (como es el caso del tópico de puntos), hacer un seguimiento de su comportamiento, para poder así determinar cómo evoluciona su comportamiento con los puntos, y ver como este servicio se puede potenciar.
- Para los otros tópicos que presentan una cantidad mayor de datos, una recomendación es la de hacer es repetir la metodología para pequeñas sub-muestras de los datos, de manera de lograr armar una estrategia corporativa más completa para cada uno de los tópicos.

Con respecto al modelo de fuga a utilizar, se tienen las siguientes consideraciones:

- Dados los resultados para la métrica de fuga, se recomienda la utilización de la métrica utilizada dentro de la memoria, con un R/F y un R mínimo.
- Con respecto al modelo a utilizar, se recomienda por ahora utilizar un modelo de fuga tradicional. Esto porque logra abordar la cartera completa de clientes y tiene mejor desempeño en ciertas métricas claves.
- Potenciar por medio de tecnologías de la empresa medios de comunicación escritos con la compañía. Esto con el objetivo de poder tener la posibilidad de generar un modelo de fuga similar al entregado por el modelo que considera las variables de minería de texto con una muestra de clientes en que todos tenían comentarios. Esto se debe a que hacen más fácil la gestión de clientes y tiene un mejor desempeño que el modelo de fuga tradicional.

8. CONCLUSIONES

De las diferentes secciones se pueden obtener ideas interesantes con respecto a la fuga y con respecto a las interacciones escritas de la empresa.

Por una parte, al mirar la cantidad de interacciones que hay dentro de cada uno de: tópicos, sentimientos y tópicos y sentimientos en conjunto, se pueden ver cuáles son los elementos más importantes desde la perspectiva del cliente. Entre los datos interesantes obtenidos de la clasificación de los diferentes comentarios, se puede ver que algunos cuentan con una importante cantidad de datos y alta concentración de comentarios negativos. Este es el caso de, por ejemplo, la tarjeta. Este tópico es el con mayor cantidad de datos, lo que lo hace el más importante desde el punto de vista del cliente si se estudia con respecto al número de comentarios que hay, pero en el que un alarmante 33% tiene comentarios negativos. Por otra parte existen tópicos que están en un mayor grado concentrados en lo positivo, como es el caso de intereses (37% de estos datos son positivos), información y comunicación (50% de estos datos son positivos) y atención (41% de estos datos son positivos). Esto da luces de cuáles son los mejores y peores aspectos de la compañía, pudiéndose formar estrategias para explotarlos.

Por medio del estudio con la herramienta de propensity score matching, fue posible determinar la existencia de dos tipos de comentarios y los impactos que estos tienen en el futuro. Por una parte, existen comentarios que, independientemente su sentimiento, llevan a un mayor gasto por parte del cliente en un futuro. Una posible causa de la existencia de este tipo de comentarios es la autoselección de clientes (quienes emiten comentarios son clientes que suelen gastar una cantidad de mayor con la tarjeta). Por otra parte se tienen los comentarios que se ven fuertemente afectados por el sentimiento asociado. Estos comentarios aumentan a futuro el gasto de ser positivos, y lo disminuyen en caso de ser negativos. Esto puede deberse a que, al ser clientes más leales a la compañía, tienen exigencias mayores que usuarios menos frecuentes. Lo importante de esta sección es que permite dar valor monetario a los tópicos y sentimientos, y en base a esto priorizar los conocimientos obtenidos de la clasificación de las interacciones. Unos ejemplos del primer tipo de comentarios son los del tópico del holding (genera un aumento de $94.355 para comentarios negativos, $107.730 para comentarios neutros y $55.484 por comentarios positivos en promedio por cliente) y el tópico de otros (genera un aumento de $50.699 para comentarios negativos, $104.540 para comentarios neutros y $50.208 por comentarios positivos en promedio por cliente). Para el ultimo tipo de impacto, el caso más representativo es el de puntos (genera una disminución de $60.097 para comentarios negativos, $45.774 para comentarios neutros y $262.263 por comentarios positivos en promedio por cliente). Este último es un caso interesante dado que presenta pocos datos.

Con respecto a los modelos de fuga resultantes, se pudieron comparar los diferentes modelos y extraer diferentes conclusiones de ellos. Por una parte, un modelo de fuga que considera las variables de minería de texto solo para clientes con comentarios tiene un desempeño superior a cualquier modelo. De no cumplir este requisito, la mejor alternativa pasaría a ser un modelo de fuga tradicional, dado que en métricas claves tiene un mejor

desempeño que los otros modelos (en el caso del modelo con toda la cartera y sin homologación la accuracy y el AUC es superior por 1,4% en ambos casos y en caso del modelo con homologación es superior por 2,2% y 1,2% respectivamente para las mismas métricas). Con esto se puede decir que se cumple el objetivo general de la memoria, pero que este tipo de modelos es solo utilizable dentro de la fuga cuando todos los clientes cuentan con un comentario.

Se concluye además que el modelo homologado se comporta de manera muy similar al modelo que considera toda la cartera con variables de minería de texto sin homologar. Esto se debe a que el modelo homologado presenta un menor accurracy que el modelo con toda la cartera y variables de minería de texto sin homologar (0,4% menos de accuracy), presenta un AUC levemente mayor (0,5% más de AUC).

9. BIBLIOGRAFÍA

[1] Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164(1), 252-268.

[2] Miglautsch, J. R. (2000). Thoughts on RFM scoring. Journal of Database Marketing & Customer Strategy Management, 8(1), 67-72.

[3] Birant, D. (2011). Data Mining Using RFM Analysis. In Knowledge-oriented applications in data mining. InTech.

[4] 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.

[5] Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval, 14(2), 178-203.

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[7] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[8] Olmos, A., & Govindasamy, P. (2015). Propensity scores: a practical introduction using R. Journal of MultiDisciplinary Evaluation, 11(25), 68-88.

[9] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[10] Goethals, O. M., & Rokach, L. (2005). The Data Mining and Knowledge Discovery Handbook.

[11] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[12] Cortez Sánchez, Víctor (2016), "Diseño e implementación de un sistema para monitorear el consumo y opinión sobre la marihuana en twitter", Memoria, Departamento de ingeniería Civil Industrial, Universidad de Chile, Chile

[13] Gallardo Mesa, Cristóbal (2016), “Identificación de clientes con patrones de alta interacción con los drivers de una tarjeta de crédito” , Memoria, Departamento de ingeniería Civil Industrial, Universidad de Chile, Chile

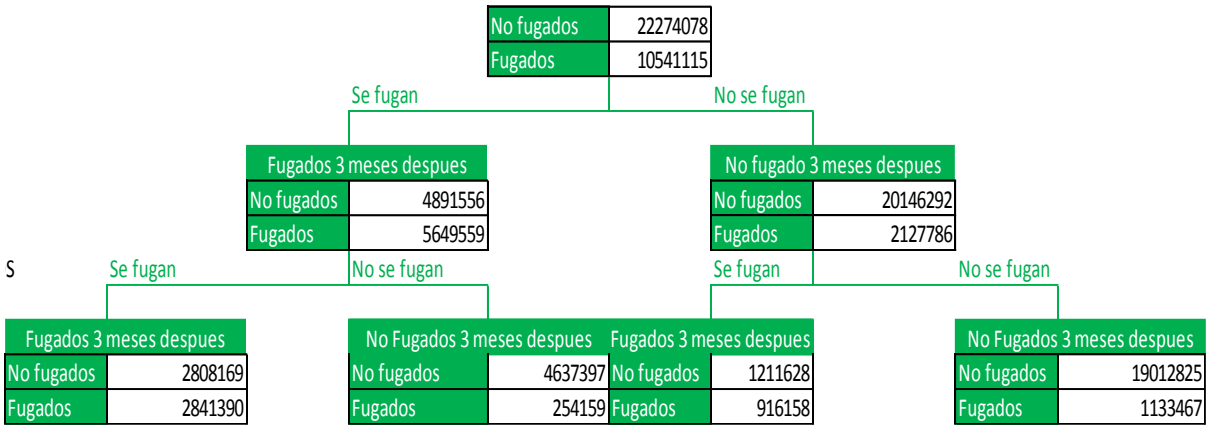
[14] Sepúlveda Jullian, Catalina (2015), “Metodología para estimar el impacto que generan las llamadas realizadas en un call center en la fuga de los clientes utilizado técnicas de text mining” , Memoria, Departamento de ingeniería Civil Industrial, Universidad de Chile, Chile

[15] Contreras Piña, Constanza Daniela (2014), “Extracción de conocimiento nuevo desde los reclamos recibidos en el servicio nacional del consumidor mediante técnicas de text mining”, Memoria, Departamento de ingeniería Civil Industrial, Universidad de Chile, Chile

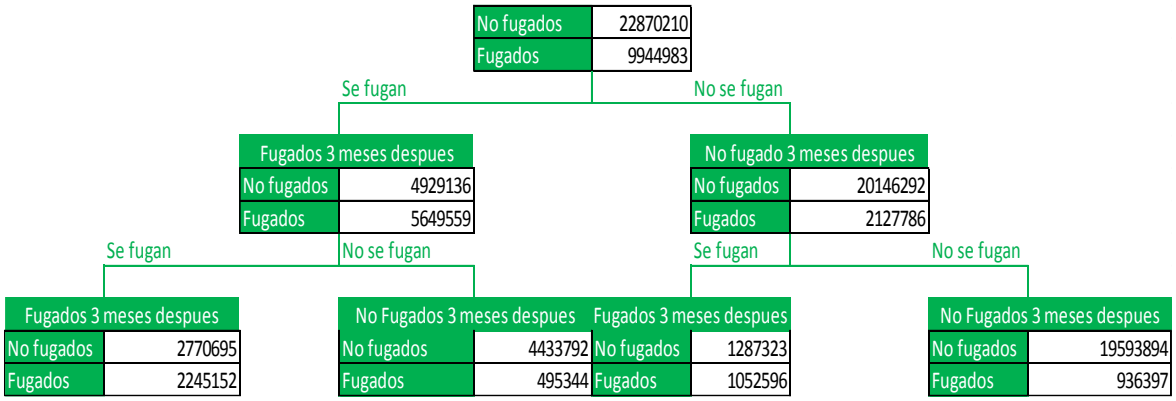
[16] Roobaert, D., Karakoulas, G., & Chawla, N. (2006). Information gain, correlation and support vector machines. *Feature extraction*, 463-470.

10.ANEXOS

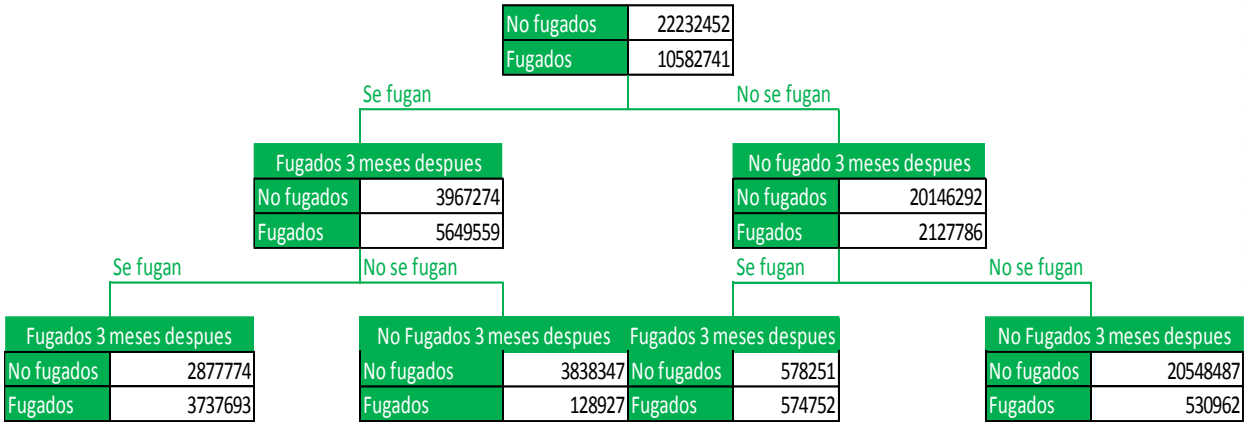
10.1 Anexo I: Árbol de evolución R/F



10.2 Anexo II: Árbol de evolución Recency



10.3 Anexo III : Árbol de evolución Frecuencia



10.4 Anexo IV: Resultados Pitman-Yor

Topico1	
compra	0,060
monto	0,039
desconocer	0,034
realizar	0,030
solicitar	0,029
cuota	0,029
transacción	0,022
internet	0,017
gracias	0,017
devolución	0,016
Intereses	

Topico2	
bueno	0,045
atención	0,042
atencin	0,022
producto	0,020
rápido	0,013
punto	0,011
excelente	0,011
poder	0,010
esperar	0,010
servicio	0,010
Atención	

Topico3	
pin	0,187
clave	0,182
entrega	0,098
contestar	0,044
multi	0,044
ivr	0,039
clav	0,028
activo	0,022
web	0,022
entregar	0,016
clave/internet	

Topico4	
código	0,191
bloqueo	0,132
bloquear	0,106
titular	0,101
bip	0,035
adicional	0,024
visa	0,023
tt	0,023
solicitar	0,022
llamar	0,017
Tarjeta	

Topico5	
actualizar	0,092
numerar	0,070
sucursal	0,051
dato	0,049
celular	0,046
derivar	0,046
validar	0,041
móvil	0,031
eecc	0,029
cambio	0,021
Sucursal/Actualización de datos	

Topico6	
pago	0,063
contar	0,061
solicitar	0,046
monto	0,044
rut	0,030
realizar	0,026
error	0,023
ingresar	0,019
devolución	0,018
gracias	0,015
Pagos y cupo	

Topico7	
falabella	0,211
seguro	0,063
compra	0,040
derivar	0,039
banco	0,035
despacho	0,034
internet	0,033
fono	0,032
tienda	0,029
garantía	0,027
Holding	

Topico8	
llamar	0,095
corto	0,053
avance	0,027
consulta	0,023
indicar	0,018
super	0,017
cupo	0,017
informar	0,017
información	0,016
contar	0,015
Información/Comunicación	

Topico9	
solicitar	0,031
punto	0,022
realizar	0,020
contar	0,017
bajo	0,016
pat	0,016
fecha	0,014
cargo	0,014
monto	0,014
gracias	0,013
Puntos	

Topico10	
contar	0,018
solicitar	0,018
hacer	0,012
solución	0,011
falabella	0,010
pago	0,009
explicar	0,009
realizar	0,009
poder	0,009
pagar	0,009
Otros	

10.5 Anexo V: Reglas comunes Modelo de fuga tradicional

Reglas de decisión principales para categoría de objetivo '1'

Regla de decisión	Categoría más frecuente	Precisión de regla	Precisión de bosque	Índice de grado de interés
(VAR_POR_MONTO_GASTO <= -0.9500903773584908) and (NUM_RUBROS_OT_3M <= 0.0)	1	0,199	0,423	0,064
(VAR_SCORE > 0.018843734939759035) and (F_PUNT_ACUM <= 793.8663101604276) and (F_PUNT_ACUM <= 1613.168367346938)	1	0,190	0,401	0,059
(VAR_SCORE > 0.1087169696969697) and (SATURACION_CUPO <= 0.07542158186340113) and (NUM_RUBROS_OT <= 0.0)	1	0,182	0,372	0,054

10.6 Anexo VI: Reglas comunes Modelo de fuga con variables de minería de texto considerando exclusivamente a clientes con comentarios

Reglas de decisión principales para categoría de objetivo '1'

Regla de decisión	Categoría más frecuente	Precisión de regla	Precisión de bosque	Índice de grado de interés
(SATURACION_CUPO <= 0.056530525817417486) and (ENGANCHADO_CMR = {0}) and (SOW_SEGUROS <= 0.7750833333333333) and (DEUDA <= 77144.13095238093)	1	0,187	0,269	0,046
(SATURACION_CUPO <= -0.02258317216745998) and (VAR_SOW <= 0.08671628571428569) and (NUM_RUBROS_OT_3M <= 0.0) and (F_PUNT_ACUM <= 2495.0521327014244) and (SATURACION_CUPO <= 0.056530525817417486)	1	0,156	0,276	0,038
(ENGANCHADO_CMR = {0}) and (SUM_MONTO_OT <= 186748.5) and (NUM_RUBROS_OT_3M <= 0.0) and (NUM_RUBROS_OT <= 1.0) and (SOW_SEGUROS_3M <= 0.10448722222222223)	1	0,176	0,220	0,037

10.7 Anexo VII: Modelo de minería de texto considerando el total de los clientes

Reglas de decisión principales para categoría de objetivo '1'

Regla de decisión	Categoría más frecuente	Precisión de regla	Precisión de bosque	Índice de grado de interés
(GRUPO_COMENTARIO = {0 - NEGATIVO,0 - NEUTRO,2 - NEUTRO,2 - POSITIVO,3 - NEUTRO,3 - POSITIVO,5 - NEUTRO, 5 - POSITIVO,6 - NEGATIVO,7 - NEUTRO,8 - NEGATIVO,8 - NEUTRO, 8 - POSITIVO,9 - NEGATIVO,9 - POSITIVO, Sin Valor}) and (VAR_SOW > 4.191094619666056E-5) and (F_PUNT_ACUM <= 1608.3216374269002) and (NUM_RUBROS_OT <= 0.0) and (SOW_RECAUDACION <= 0.005271818181818181)	1	0,159	0,352	0,045
(SATURACION_CUPO <= -0.07924879791020906) and (NUM_RUBROS_OT <= 0.0) and (SOW_RECAUDACION <= 0.005271818181818181) and (SATURACION_CUPO <= 0.07521193257495118)	1	0,180	0,272	0,044
(SATURACION_CUPO <= -0.07924879791020906) and (SUM_MONTO_OT <= 187019.10240963858) and (SOW_TIENDAS_DPTO <= 0.6843083333333333) and (NUM_RUBROS_OT <= 0.0)	1	0,171	0,282	0,043

10.8 Anexo VIII: Reglas comunes Modelo de fuga con variables de minería de texto generadas con logit multivariado de homologación de clientes

Reglas de decisión principales para categoría de objetivo '1'

Regla de decisión	Categoría más frecuente	Precisión de regla	Precisión de bosque	Índice de grado de interés
(DELTA_NUMERO_RUBROS_OT <= 0.0) and (F_PUNT_ACUM <= 694.351612903226) and (SATURACION_CUPO <= 0.0543083947368421 1) and (F_PUNT_ACUM <= 1631.6981132075475)	1	0,189	0,228	0,041
(SOW_TIENDAS_DPTO_3M <= 0.38181176470588235) and (SATURACION_CUPO <= 0.15042796428571426) and (SOW_SEGUROS_3M <= 0.25005875) and (NUM_RUBROS_OT <= 0.0) and (SUM_MONTO_OT_3M <= 190340.80790960457)	1	0,179	0,193	0,034
(SATURACION_CUPO <= 0.0543083947368421 1) and (SOW_COMUNICACIONES_3M <= 0.0) and (SOW_SEGUROS_3M <= 0.25005875) and (DELTA_NUMERO_RUBROS_OT <= 0.0) and (SOW_RECAUDACION <= 0.0)	1	0,162	0,220	0,033