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

ESTUDIO DE LA SATURACIÓN EN EMAIL MARKETING PARA UN NEGOCIO DE RETAIL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MIGUEL ALEJANDRO GUTIÉRREZ CAMPOS

PROFESOR GUÍA:
CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
DANIEL SCHWARTZ PERLROTH

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Miguel Alejandro Gutiérrez Campos
FECHA: 01/04/2019
PROFESOR GUÍA: Carolina Segovia Riquelme

ESTUDIO DE LA SATURACIÓN EN *EMAIL MARKETING* EN UN NEGOCIO DE *RETAIL*

A nivel mundial, el *email marketing* ha mostrado ser una herramienta muy utilizada por las empresas para promocionar sus productos y retener a sus clientes. Por esto, se envían campañas masivas esperando generar en el cliente una eventual compra. Sin embargo, esta facilidad con que se puede enviar un *email* puede ser contraproducente, dado que existen ciertas prácticas que pueden generar un impacto negativo sobre ciertos clientes. El presente trabajo busca abordar esta problemática, teniendo como hipótesis que la baja en las características de apertura de los clientes puede provocar cambios negativos en su comportamiento de compra, que es lo que se va a denominar saturación en este contexto.

Los resultados muestran que los clientes con tasas de apertura superiores al 50% y cuyo grado de adopción *web*, que es el porcentaje de transacciones realizadas a través del canal *web*, es de al menos un 80%, presentan en promedio caídas desde un 22% en sus transacciones cuando tienen una caída en su tasa de apertura de un 60% durante 4 semanas. Se observa que los clientes que usan un computador para revisar sus correos son un 42% más propensos a saturarse que los que no lo hacen. Por su parte, enviar un *email* durante un domingo los hace más propensos en un 81% respecto a no hacerlo. También se observa que enviar *emails* desde las 19 hrs. hasta las 9 hrs. disminuye la propensión a saturarse en un 12%.

Como recomendación para la empresa se propone estudiar con mayor detalle los potenciales perjuicios de enviar de *emails* durante los días domingo e identificar con información histórica a los clientes que usen su computador para abrir *emails* y entender así si son realmente más propensos a saturarse. También se recomienda entender los posibles beneficios de enviar *emails* fuera del horario de trabajo, estudiando distintos rangos horarios incluidos dentro de las 19 hrs. a las 9 hrs. Como trabajo futuro se recomienda estudiar variables que no se hayan controlado en este trabajo. Particularmente, eventos especiales como los *cyber*, ya que concitan un interés particular sobre los clientes que, probablemente, no sea el mismo que cualquier otra campaña común. Se sugiere también realizar un análisis basado en un diseño experimental variando la cantidad de toques.

Si bien se observa que los efectos por saturación en *email marketing* sólo aplican a un 6,64% del segmento con mayor tasa de apertura, se prevé que haya un aumento en el uso del canal *web* para comprar, por lo que es válido anticipar esta situación que se vislumbra como un problema real en el futuro.

AGRADECIMIENTOS

“… pasó mucho tiempo contemplando antes de llegar a una respuesta. La respuesta fue gratitud.”

A la empresa donde desarrollé este trabajo por la oportunidad brindada. Gracias.

A los profesores de la comisión, Carolina y Pablo, por guiarme y aconsejarme en este último trabajo. Mención para el profesor Daniel, que se incorporó al final y fue muy gratificante ya que fue uno de los profesores que más admiré durante mi paso por la U.

Volviendo el tiempo atrás, recuerdo a todos quienes estuvieron en mi desarrollo académico y personal. En ocasiones puede ser injusto mencionar a alguien, ya que se deja de mencionar al resto.

A mi colegio Padre Hurtado de Pudahuel, donde inicié mi tránsito académico y viví muchas experiencias. Mención especial al tío Claudio, que siempre creyó en mí.

Al Instituto Nacional, por ser el lugar donde conocí a mis mejores amigos y queridos profesores. Mención para el profesor Rafael Terreros, que nos supo guiar con gran dedicación en un mundo totalmente nuevo para niños de 13 años.

A mis amigos de la U que se concentran principalmente en los de la sección 5 del 2013 y mis amigos de industrias. Gracias por los innumerables carretes, pichangas, estudios y lindos momentos en general. Los quiero a pesar de todo.

A mis vecinos, por siempre ayudarme cuando lo necesité durante este tiempo. Gracias sobre todo a mis queridos vecinos, Mauricio y Marcela, que fueron como padres postizos.

A mi Nachita, por acompañarme durante todo este proceso final en la Universidad con tanto amor y paciencia.

Finalmente, a mi familia, que han sido mi soporte desde siempre. Tías, tíos, primos, abuelita. Frodo, mi perrito querido, me llenó de alegría al llegar de un día agotador. Cami, hermanita, gracias por ser como eres. Papá, gracias por darnos todo sin esperar nada a cambio.

Por último, mamá. Sin duda que sin ti esto no hubiera ocurrido. Gracias por siempre buscar lo mejor para nosotros. No me alcanzará la vida para devolver todo lo que me diste.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

I. INTRODUCCIÓN

El presente trabajo busca estudiar la saturación, entendiendo las causas que lo provocan y buscando consecuencias en términos transaccionales para la empresa, con el fin de ofrecer recomendaciones que eviten este problema. Se propone un sistema de costos en los envíos de *emails* que se obtenga en función de la probabilidad de saturación y el impacto transaccional que pueda tener como solución a este problema.

La saturación con *emails* y el mal uso de este canal, manifestado a través del *spam*, disminuye las tasas de apertura según se ha alertado (Priore, 2000). También se ha establecido la presencia de saturación en otros medios de *marketing*, como el *SMS* (Gauzente, Ranchhod, Guraud; 2008) donde los resultados indican que el género de los encuestados, la frecuencia de uso y la duración del uso influyen significativamente en el nivel de saturación específico del remitente, lo que crea posibles variables de segmentación para los usuarios de teléfonos móviles. Además, se debe tener en cuenta que proporcionar incentivos excesivos a los clientes para recomendar productos podría ser contraproducente al debilitar la credibilidad de estos (Leskovec, Adamic, Huberman; 2007).

En capítulo II se define el objetivo general y los objetivos específicos que establecen la línea de trabajo del proyecto. En el capítulo III se presenta el marco teórico con los conceptos y métodos existentes en la literatura que se pueden utilizar para abordar el problema.

En el capítulo IV se presenta la metodología elegida para el desarrollo del trabajo en base a los capítulos anteriores. En el capítulo V se presentan los alcances y resultados esperados del trabajo. En el capítulo VI se exhibe el desarrollo metodológico y los resultados obtenidos.

En el capítulo VII se exhiben las conclusiones del trabajo, junto con una breve reflexión sobre las limitaciones del proyecto y propuestas para trabajos futuros. Finalmente, en el capítulo VIII se muestran algunos conceptos relevantes que pueden desconocerse a priori.

1.1 Empresa

La empresa es una tienda por departamento que cuenta con operaciones en Chile,

Argentina, Perú, Colombia y Brasil. Pertenece a un *holding* que posee negocios en el sector de *retail*, financiero, mejoramiento del hogar, supermercados y negocio inmobiliario.

La empresa ofrece productos para uso personal y del hogar, ordenados por las categorías vestuario y calzado, artículos de belleza, artículos electrónicos y electrodomésticos, muebles y accesorios de decoración. Cuenta con marcas propias, marcas internacionales, marcas locales y de segunda generación.

Posee tiendas especializadas en vestuario y calzado de marcas exclusivas y de segunda generación, donde se ofrece una mayor variedad de productos de esas marcas que en las grandes tiendas.

A continuación, se muestra una tabla con información relativa a la presencia de la empresa en la región:

País	Participación de mercado	Número de tiendas	Superficie de ventas [m2]
Chile	23%	45	318.333
Perú	18%	29	176.962
Colombia	7%	26	174.831
Argentina	2%	11	58.426

Tabla 1. Presencia de la empresa en Latinoamérica.
Fuente: Memoria anual 2017 de la empresa.

Durante el año 2018, se identifican 5.879.665 clientes. Se clasifican en clientes normales y clientes de alto valor. Estos últimos son un 10% aproximadamente. Estos clientes cumplen con ciertos criterios de gasto y visitas durante el período de un año.

La distribución de los clientes según su nivel socioeconómico se concentra principalmente en el nivel medio de la población, representado por el sector D con un 45% del total. Esto se observa en la siguiente ilustración:

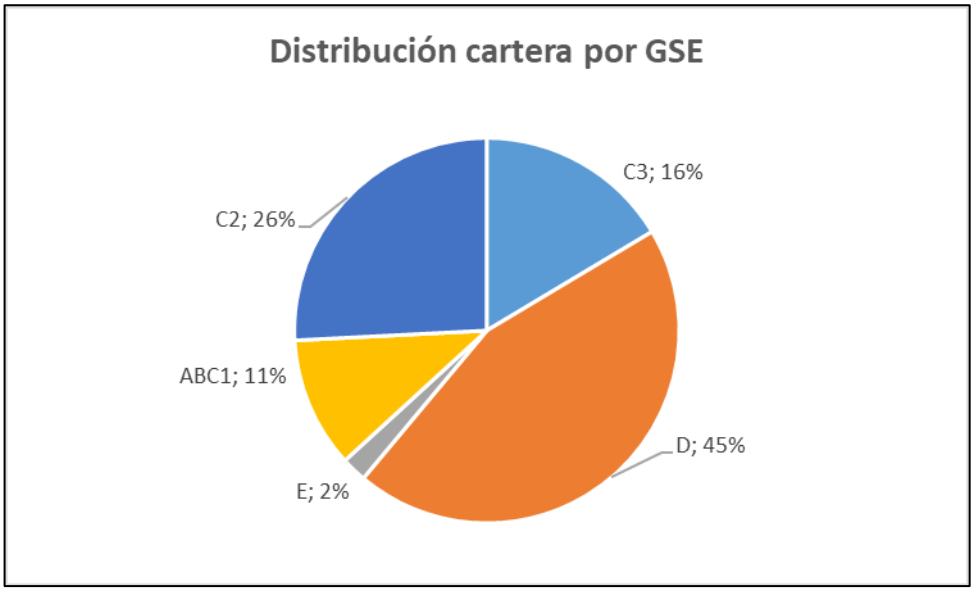

Ilustración 1. Distribución de la cartera según GSE
Fuente: Elaboración propia con datos de la empresa.

El porcentaje de personas en el sector ABC1 se condice con el porcentaje de personas de alto valor para la empresa, dado que es un sector con mayor potencial de gasto.

1.2 *Email marketing*

Actualmente, el correo electrónico es una herramienta ampliamente usada en el mundo. Más de un 34% de las personas en el planeta usan un correo, lo que corresponde aproximadamente a 2,5 billones de individuos. Se prevé que esta cifra aumente a 2,8 billones en los próximos dos años. Además, se estima que se envían 196 billones de *emails* diariamente, de los cuales 109 billones, que corresponden a un 55% aproximadamente, son *emails* comerciales[1].

Por lo anterior es que las empresas consideran tan importante el uso del *email* como canal de *marketing*, ya que pueden llegar a gran cantidad de gente. Además de esto, existen razones específicas por las cuales el *email marketing* es relevante, como se enumera a continuación[2]:

[1] 2018. *Email* Statistics Report, 2014-2018. [En línea]. Palo Alto. Sara Radicati. < http://www.radicati.com/wp/wp-content/uploads/2014/01/*Email*-Statistics-Report-2014-2018-Executive-Summary.pdf> [07-02-2019]
[2] 2018. 6 Reasons Why *Email Marketing* Is Important For Your Internet *Marketing*. [En línea]. < https://inboundrocket.co/blog/6-reasons-why-*email-marketing*-is-important-for-internet-*marketing*> [07-02-2019]

- Más efectivo que las redes sociales (para adquisición de clientes): si bien las redes sociales son un componente muy importante en cualquier estrategia de *marketing*, se tiene que la adquisición de clientes a través de *email marketing* es superior a la que se alcanza a través de otros medios como las redes sociales o la búsqueda pagada. A continuación, el resultado de una encuesta a profesionales ligados a la industria del *retail* en Estados Unidos realizada por eMarketer.com:

Medio	Adquisición	Retención
Email marketing	81%	80%
Búsqueda orgánica	62%	36%
Búsqueda pagada	59%	43%
Redes sociales	51%	44%
Marketplaces	15%	11%

Tabla 2. Resultados de encuesta sobre adquisición y retención de clientes por distintos canales.
Fuente: www.eMarketer.com[3]

La tabla 1 muestra que el 81% de las personas entrevistadas creen que el *email marketing* logra adquirir nuevos clientes, cifra que supera el resto de los medios de la tabla.

- Económico: permite llegar a una gran cantidad de personas y el costo por cada *email* enviado es despreciable. De hecho, es el canal de *marketing* que genera el mayor ROI (retorno a la inversión). Se tiene que por cada 1 USD$ se generan 40 USD$ de retorno[4]. El siguiente gráfico ilustra esta realidad:

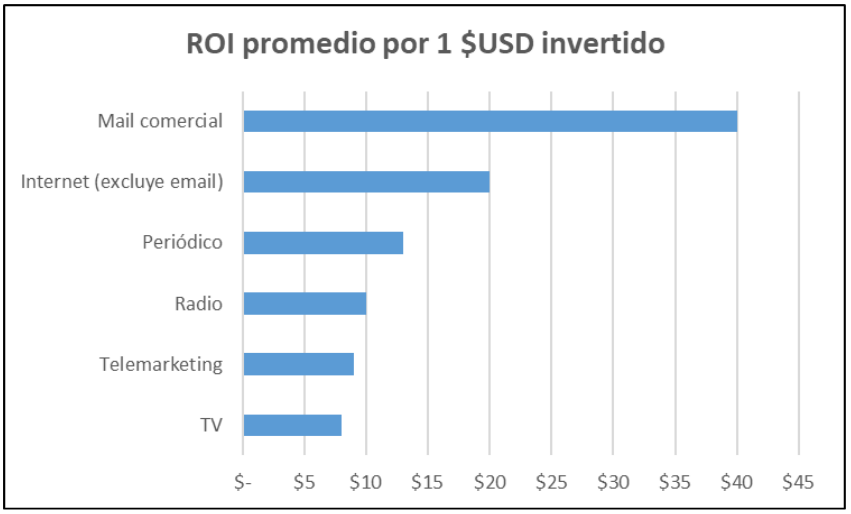

Ilustración 2. ROI para distintos canales de *marketing*.
Fuente: Direct *Marketing* Association.

[3] Digital Tactics that Drive Customer Acquisition vs. Retention According to US SMB *Retail* Professionals, Marzo 2016. [08-02-2019]
[4] Direct *Marketing* Association. 2014. Average Roi per 1$ in ad spending. [08-02-2019]

- Orientados a la acción: las campañas comerciales buscan incidir en el comportamiento de las personas, ya sea logrando que visiten la página *web* de la empresa o bien logrando que la persona compre algún producto.
- Medibles: se puede conocer, a través de softwares de *email marketing*, si la persona abre o no el *email*, si hizo algún click o si se desuscribe. Con esto se puede conocer cómo se desempeña una campaña.

Dada la importancia de este canal de comunicación, el *holding* al cual pertenece el negocio cuenta con reglas de gestión de este canal con el fin de dar un buen uso a la herramienta. Esto es lo que se conoce como la política de toques del *holding*. Esta política consiste en restricciones a la cantidad de envíos que cada negocio del *holding* debe realizar. Estas restricciones aplican a la cantidad mínima y máxima de envíos de correos.

Para esto, se agrupa a los negocios en dos grupos, uno de alto nivel de envío y otro de bajo nivel. Además, se agrupa a las personas en cuatro grupos, los cuales son de alta y media-alta apertura, baja y media-baja apertura, pocos envíos junto a sin envíos y finalmente el grupo de no apertura. El negocio estudiado corresponde al grupo de nivel alto de envíos. La ilustración 3 esquematiza esta idea.

Allí se observa que para los negocios con un alto nivel de envíos la mínima cantidad de *emails* que pueden enviar es uno por semana. La máxima cantidad son cinco para el grupo de alta y media-alta apertura y tres para los grupos de baja y media-baja apertura, pocos envíos y sin envíos. El grupo de no apertura escapa a estas reglas, ya que sólo recibe *triggers*.

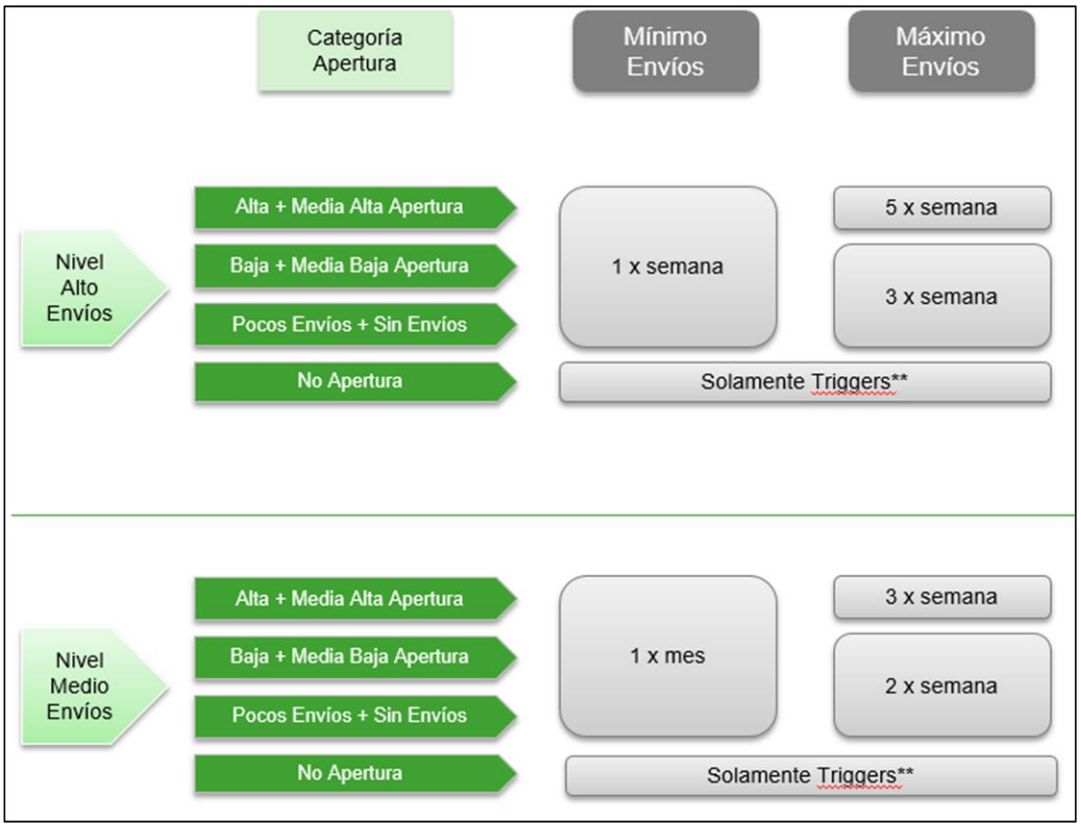

Ilustración 3. Política de toques del *holding*.
Fuente: Información del *holding*.

1.3 Justificación del proyecto

El *email marketing* es un canal de comunicación ampliamente usado por las empresas. Sus beneficios se presentan en la sección de Motivación. La empresa es consciente de esto y es así como el año 2016 desarrolla un total de 882 campañas de *email marketing* que se traducen en 58 millones de envíos. El desglose de estas campañas se observa en la tabla 3.

Tipos de campaña	Cantidad de campañas	Cantidad de clientes [K[5]]
Apoyo masivo	221	42.616
Proveedor belleza	82	2.081
Stand alone	47	1.522

[5] K equivale a mil.

Meta compra	7	1.304
Retención	41	2.040
Reach	62	3.667
Trigger hitos	71	186
Trigger venta cruzada	79	490
Trigger recompra	10	44
Trigger navegación	99	449
Trigger garantía extendida	6	292
Retención F	9	53
Reach 2.0	25	250
Overlap Cliente F – P/E	4	373
Cupón POS	119	1.999
Total 2016	**882**	**58.093**

Tabla 3. Resumen envío *emails* de 2016.
Fuente: Información del negocio.

La idea de estas campañas es presentar nuevos productos o promociones, buscando ser un *driver* para aumentar las ventas. Dado esto, es un problema cuando las personas que usualmente leen los *emails* de la empresa dejan de hacerlo, pues no se entrega el mensaje deseado.

Se han observado casos donde clientes con una cierta tasa de apertura[6] tienen bajas abruptas en este indicador, tal como se observa en la próxima ilustración. Así es como nace el concepto de saturación en *email marketing*. La palabra saturación se define como “colmar, llenar de modo que exceda”[7].

[6] La tasa de apertura corresponde al número de *emails* abiertos dividido por el número de *emails* recibidos en un espacio temporal.
[7] Real Academia Española. (2001). *Diccionario de la lengua española* (22.ªed.). En línea <http://www.rae.es/rae.html> [08-02-2019]

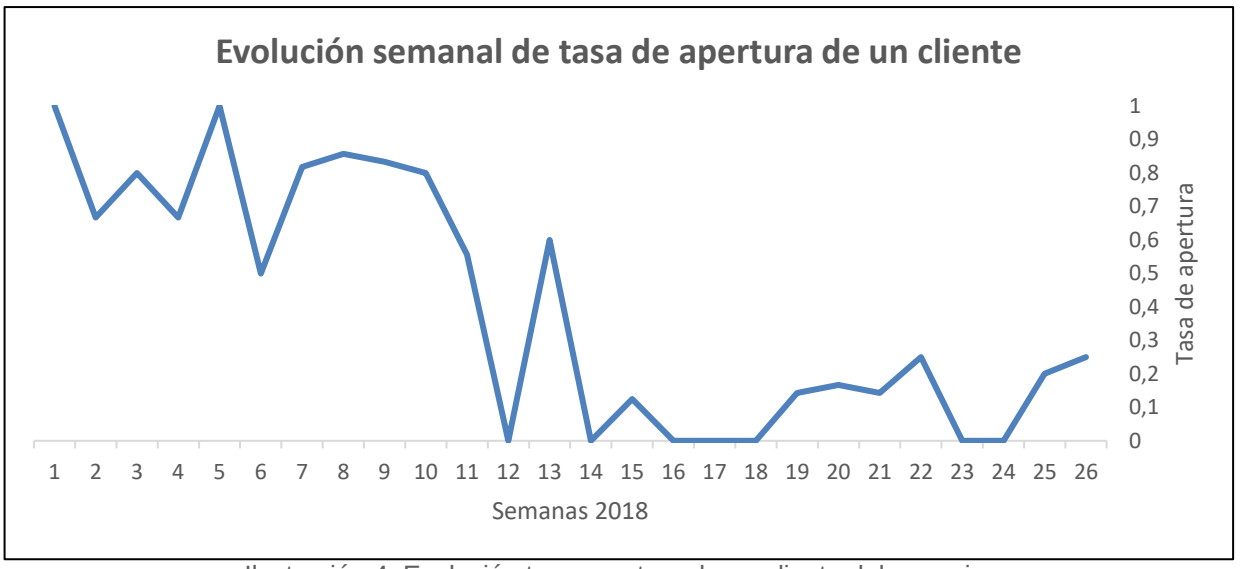

Ilustración 4. Evolución tasa apertura de un cliente del negocio.
Fuente: Elaboración propia con datos de la empresa.

- Tasa de referencia: corresponde a la tasa de apertura que tiene un cliente antes del instante en que presenta una caída en la tasa.
- Caída en la tasa de apertura: corresponde a una disminución en la tasa de apertura producida a partir de un instante del tiempo.
- Temporalidad: corresponde a la duración de la caída en la tasa de apertura.

Un fenómeno se define como la unión de estos tres conceptos. Por ejemplo, un fenómeno puede ser el caer en la tasa de apertura un 50% respecto a su tasa de referencia por dos semanas.

La siguiente ilustración muestra los conceptos de manera gráfica:

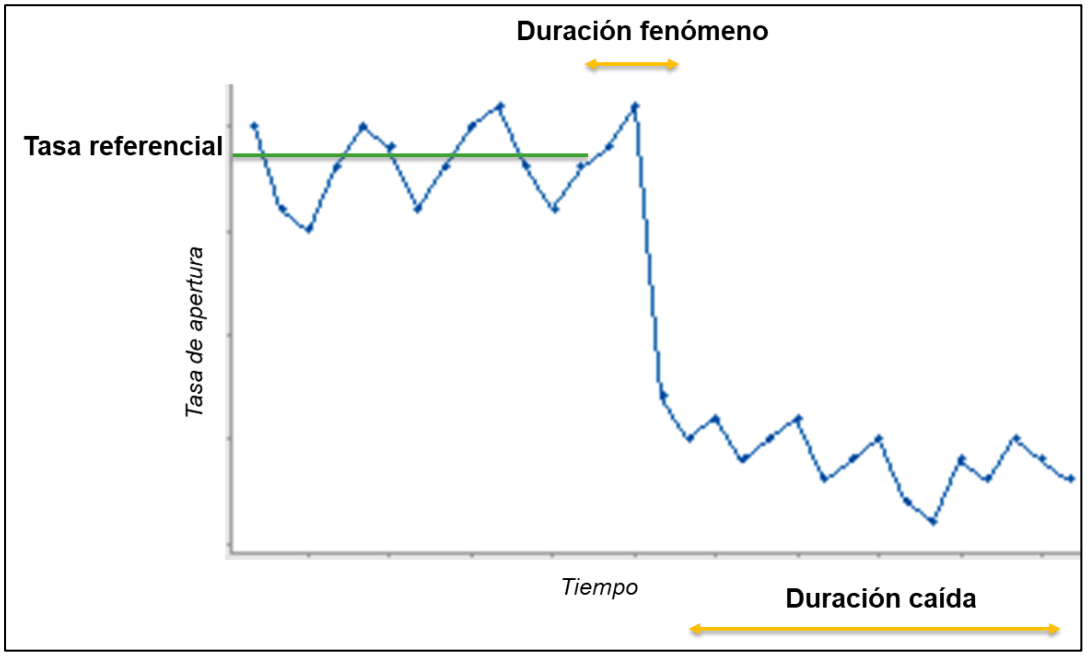

Ilustración 5. Esquema de ejemplo de componentes de saturación.
Fuente: Elaboración propia con datos ficticios.

Un ejemplo numérico para simplificar la comprensión de esta ilustración: un cliente tiene una tasa referencial durante un mes de un 80% en promedio. Luego, pasan dos semanas, que es lo que dura el fenómeno, donde posterior a esto la nueva tasa de apertura es de un 20%. Es decir, tuvo una caída de un 75%[8].

La hipótesis de investigación es que la baja en las características de apertura en los clientes puede provocar cambios negativos en su comportamiento de compra.

[8] Notar que todos estos números sólo son hipotéticos.

II. OBJETIVOS

2.1 Objetivo general

Recomendar reglas de gestión de *email marketing* a un negocio de *retail* para mitigar las consecuencias de la saturación por envíos de *emails* y así evitar un descenso en las características transaccionales de sus clientes.

2.2 Objetivos específicos

- Segmentar a los clientes en función de la cantidad de envíos y tasa de apertura.
- Definir el criterio de saturación para los distintos segmentos en función del impacto transaccional sobre los clientes.
- Calcular la probabilidad de saturación de los segmentos por día.
- Proponer el cálculo de un costo de envío para los *emails*, en función de la probabilidad de saturación y el impacto transaccional.

III. MARCO TEÓRICO

El presente capítulo pretende introducir los conceptos relevantes para el desarrollo del trabajo.

En vista del primer objetivo, en 3.1 se presenta la técnica usada para realizar segmentaciones; en 3.2 se analizan conceptos como las diferencias en diferencias, la regresión lineal simple y el *test* de hipótesis. Luego, en 3.3 se muestra el método para seleccionar variables, el tipo de modelo que permite calcular la probabilidad, cómo abordar potenciales problemas de multicolinealidad y la manera escogida para evaluar la capacidad predictiva del modelo. Con respecto al cuarto objetivo no hay un marco teórico asociado dado que no hay precedentes de esta práctica.

3.1 Segmentación

Un segmento se define como un trozo o parte cortada o separada de una cosa[9]. Segmentar corresponde a separar o dividir una cosa en segmentos, es decir, en partes.

En *marketing*, lo que se busca con la segmentación es capturar las variadas respuestas que el cliente muestra con respecto a sus características, que puede ser un hábito de compra, por ejemplo. Se busca tomar decisiones orientadas a la heterogeneidad que se presenta, pero sin tener que diseñar una estrategia de *marketing* para cada cliente.

Por eso se genera una segmentación que contenga grupos representativos de clientes. En este caso, el método usado para segmentar es el método de *k-medias*.

i. *K-medias*

K-medias es un método de agrupamiento que tiene como objetivo la partición de una población N-dimensional en *k* grupos en el que cada observación pertenece al grupo cuyo centro es más cercano. El centro de cada *cluster* es llamado también centroide.

[9] Real Academia Española. (2001). *Diccionario de la lengua española* (22.ªed.). En línea <http://www.rae.es/rae.html> [08-02-2019]

En otras palabras, se busca encontrar k vectores $c_1, \dots c_k$ y entregar una asignación a algún *cluster* $y_i \in \{1,\dots,m\}$ a cada punto x_i en el conjunto. El algoritmo de este método está basado en una aproximación entrelazada, donde la asignación a los *cluster*s y_i se establece dados los centros y los centros se computan dadas las asignaciones. Su formulación es como sigue:

$$\min_{y_1, \dots y_m, c_1, \dots c_k} \sum_{j=1}^{k} \sum_{y_i = j} \|x_i - c_j\|^2 \tag{1}$$

Asumiendo que $c_1, \dots c_k$ están dados por la iteración anterior, se tiene:

$$y_i = argmin_j \|x_i - c_j\|^2 \tag{2}$$

Posteriormente, asumiendo que las asignaciones a los *cluster*s $y_1, \dots y_m$ es dada, entonces para cualquier conjunto $S \subseteq \{1, \dots, \mathrm{m}\}$ se tiene que:

$$\frac{1}{|S|} \sum_{j \in S} x_j = argmin_c \sum_{j \in S} \|x_j - c\|^2 \tag{3}$$

Así es como dados los conjuntos de centros calculados en primera instancia, se calculan las nuevas asignaciones por el centro más cercano a cada punto x_i, y luego dadas las asignaciones actualizadas los nuevos centros se estiman tomando la media de cada *cluster*.

Con las asignaciones listas se puede mostrar el concepto de inercia, que corresponde a la suma de las distancias al cuadrado de cada elemento del *cluster* a su centro:

$$Inercia = \sum_{i=1}^{m} \sum_{x \in S_i} \|x - c_i\|^2 \tag{4}$$

Este método es fácilmente programable y computacionalmente económico. Se puede aplicar en agrupamiento por similitud, predicción no lineal, aproximación de

distribuciones multivariadas y en *test*s de independencia no paramétricos entre variables.

3.2 Definición criterio saturación

En esta parte se requiere conocer la herramienta de diferencias en diferencias, las regresiones lineales y el *test* de hipótesis.

3.2.1 Diferencias en diferencias

La herramienta de diferencias en diferencias es un método de estimación de inferencia causal utilizado ampliamente en estudios observacionales, pero que en este caso se usa con información histórica (Campos, 2016).

Se tiene al grupo de los tratados, llamado grupo tratamiento, y al grupo de los no tratados o grupo control. Se calcula el efecto del tratamiento a través de la diferencia promedio en el grupo tratamiento restada con la diferencia promedio para el grupo control. Se asume que los grupos mantienen una tendencia en la variable de respuesta en el tiempo.

Para la formulación se tiene que Y_t corresponde al valor de la variable respuesta sobre el grupo tratado y que Y_c corresponde al valor de la variable respuesta sobre el grupo control. P corresponde al período: vale 1 si se refiere al período posterior al tratamiento y 0 si se refiere al período previo. El efecto del tratamiento se calcula como sigue:

$$\text{Efecto tratamiento} = [(Y_t|\, P = 1) - (Y_t|\, P = 0)] - [(Y_c|\, P = 1) - (Y_c|\, P = 0)] \quad (5)$$

En la Ilustración 6, el efecto del tratamiento corresponde a la línea punteada vertical indicada por *Intervention effect*.

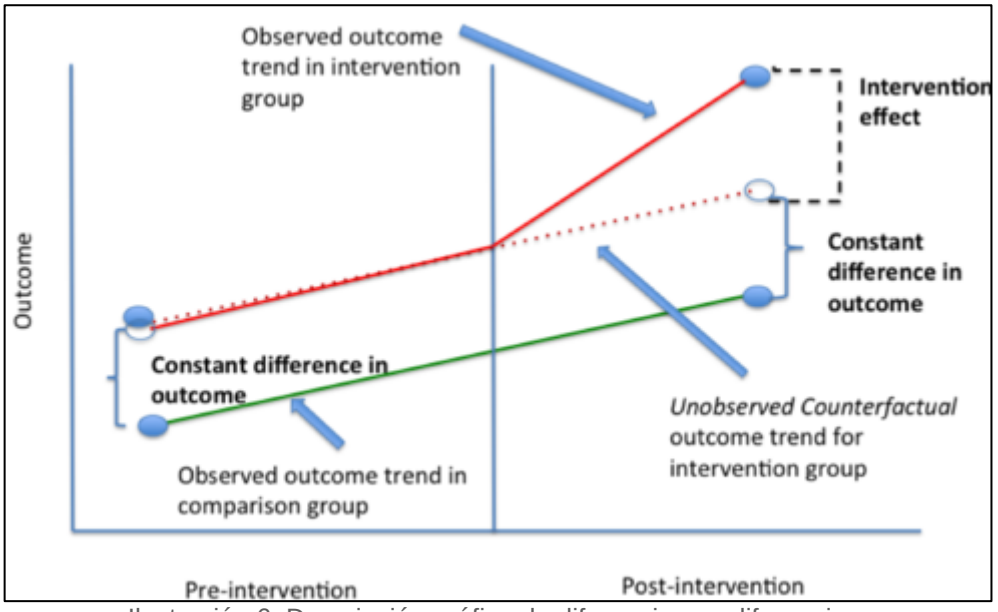

Ilustración 6. Descripción gráfica de diferencias en diferencias.
Fuente: Mailman school of public health, Columbia.

3.2.2 Regresión lineal simple

En estadística, la regresión lineal modela la relación entre una variable dependiente con una o más variables independientes. Cuando la variable dependiente es una, entonces se llama regresión lineal simple. De ser más de una, se llama regresión lineal múltiple (Freedman, 2009).

Cuando la relación entre la variable dependiente con las independientes se modela usando funciones lineales entonces se tiene un modelo lineal. Esta regresión tiene distintos usos prácticos tales como predecir o explicar. En esta sección, se busca explicar si la variación en la variable dependiente puede ser atribuida a la variación en la variable independiente. Para esto, se requiere el concepto de *test* de hipótesis y p-valor.

3.2.3 *Test* de hipótesis y p-valor

El *test* de hipótesis es una técnica de inferencia estadística que permite comprobar si la información dada por una muestra concuerda con la hipótesis estadística formulada sobre el modelo de probabilidad en estudio, de manera de establecer si se puede aceptar la hipótesis formulada[10].

[10] Departamento de matemáticas de la Universidad de la Coruña. Contraste o *test* de hipótesis. Definiciones. 2006. [10-02-2019]

Una hipótesis estadística es una suposición sobre alguna característica de un modelo de probabilidad. La hipótesis que se prueba se llama hipótesis nula, denotada por H0. La hipótesis complementaria, que se asume correcta si la hipótesis nula es rechazada es la llamada hipótesis alternativa y se denota por H1. Cuando se realiza un *test* de hipótesis existe la posibilidad de cometer errores de dos tipos:

- Error tipo I: se rechaza la hipótesis nula H0 cuando es cierta.
- Error tipo II: se acepta la hipótesis nula H0 cuando es falsa.

La siguiente tabla ilustra los tipos de errores:

		Realidad	
		H0 es Verdadero	H0 es Falso
Decisión	Rechazo H0	**Error tipo I**	Decisión correcta
	No rechazo H0	Decisión correcta	**Error tipo II**

Tabla 4. Error tipo I y error tipo II.
Fuente: Elaboración propia.

El nivel de significación, denotado por α, es la probabilidad de cometer un error tipo I. Esta probabilidad es la que se conoce como p-valor. Cuando se fija este nivel se establece la probabilidad máxima que se puede asumir de rechazar la hipótesis nula cuando en realidad es cierta. En general, se toma un α=0,05.

3.3 Modelo predictivo

Para poder obtener una probabilidad de saturación se genera un modelo probabilístico. Considerando que existe un desbalance de los datos con respecto a la variable dependiente, se utiliza la estrategia de remuestreo. Otra situación a

abordar tiene que ver con la gran cantidad de variables con las que se cuenta. Por lo tanto, se aplica el método SCAD, buscando un equilibrio entre bondad de ajuste y cantidad razonable de variables.

Para comparar los distintos modelos se usan métricas de ajuste, donde se escoge usar AIC. Luego, se formaliza un modelo con las variables restantes usando una regresión lineal logística binaria, que se destaca por la posibilidad de interpretar las variables que contiene. Finalmente, se requiere usar métricas de desempeño de la capacidad predictiva del modelo. Para esto, se utiliza la curva ROC.

3.3.1 Selección de variables

Muchas veces se dispone de una gran cantidad de posibles variables explicativas. En ese momento se debe decidir si incluir todas las variables o tomar un subconjunto de ellas. Usar muchas variables en un modelo aporta al ajuste de los datos, pero entrega una mayor varianza, mientras que incluir menos variables de las necesarias reduce la varianza, pero aumenta los sesgos.

Por tanto, el uso de un método de selección de variables busca un buen ajuste a los datos y un equilibrio entre bondad de ajuste y sencillez (González, Aneiros; 2014). Existe una clasificación para los métodos de selección de variables. Se tiene la categoría de algoritmos de selección, donde se elige el mejor modelo de forma secuencial al incluir o excluir alguna variable explicativa y los métodos de mínimos cuadrados penalizados, cuyo uso se recomienda cuando se trabaja con una gran cantidad de variables.

En esta sección se incluye un representante de cada categoría. Por el lado de los algoritmos se incluye el método *backward* y por el lado de los métodos de mínimos cuadrados penalizados el método SCAD.

i. Método SCAD

Este método pertenece a la categoría donde la idea clave es la penalización. Se prefiere por sobre el método Lasso, ya que en general selecciona una menor cantidad de variables, lo que es bastante conveniente para no complejizar el modelo en exceso. Con el método SCAD se busca evitar el sobreajuste debido al gran número de variables dependientes a través de la penalización por variables, lo que obliga a que algunas componentes del vector de parámetros β sea cero. La siguiente expresión representa a este método:

$$p_{\lambda}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j|, & si\ \ 0 \le |\beta_j| \le \lambda \\ -\dfrac{\beta_j^2 - 2a|\beta_j| + \lambda^2}{2(a-1)}, & si\ \ \lambda \le |\beta_j| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2}, & si\ \ |\beta_j| \ge a\lambda \end{cases} \tag{6}$$

$$donde\ a > 2\ y\ \lambda > 0$$

El estimador SCAD se define entonces como el que minimiza la siguiente expresión:

$$\sum_{i=1}^{n} (y_i - x'_i\beta)^2 + \lambda \sum_{j=1}^{p} p_{\lambda}^{SCAD}(\beta_j) \tag{7}$$

El parámetro λ se selecciona a través del método de validación cruzada generalizada, que se caracteriza por ser más selectivo con las variables que selecciona.

ii. Método *backward*

Se parte con un modelo complejo y en cada etapa se elimina la variable menos influyente, hasta encontrar el mejor modelo de acuerdo a alguna métrica. Se tiene entonces un criterio de significación, donde la influencia de la variable se evalúa en función de la significancia estadística de la misma; además, un criterio global donde se escoge el mejor modelo en relación a la verosimilitud que éste tenga, es decir, qué tan bien se ajusta a los datos. Para esto se usan las métricas de desempeño.

3.3.2 Métrica de ajuste

Se requiere usar una métrica de ajuste para escoger al modelo al momento de estar sacando variables. La métrica de desempeño usada es el AIC.

i. AIC

El criterio de información de Akaike (AIC) es una métrica de calidad de ajuste que permite comparar modelos probabilísticos, pero por sí solo no aporta información. Mientras menor sea el valor de este criterio, mejor para un modelo respecto de otro con mayor valor. La fórmula general para calcular el valor del AIC es la siguiente:

$$AIC = 2k - 2\ln(L) \tag{8}$$

En la fórmula, *k* corresponde al número de parámetros y L al máximo valor de la función de verosimilitud.

3.3.3 Balanceo de datos

Dado que se cuenta con datos desbalanceados, es decir, existen pocos casos positivos (situación saturada) en relación a los casos negativos (situación no saturada), se usa la estrategia de remuestreo. Esto puede consistir en agregar casos ficticios o duplicados a la situación minoritaria o bien submuestrear la clase mayoritaria[11]. Se observa en estudios donde se utiliza esta técnica que la mejor opción es submuestrear la clase mayoritaria, usando una muestra con un 60% del caso mayoritario y 40% del caso menos frecuente (Covarrubias, 2012).

3.3.4 Detección de multicolinealidad

Dada la naturaleza de las variables se examina la presencia de multicolinealidad. Esto corresponde a una correlación entre las variables explicativas de un modelo. Para detectarla se aplica el factor de intensidad de la varianza (popularmente como VIF, que es su sigla en inglés).

i. VIF

El VIF es una medida de la correlación entre las variables explicativas. Se calcula para cada variable explicativa usando la siguiente fórmula:

[11] KARAGIANNOPOULOS, M., ANYFANTIS, D., KOTSIANTIS, S. y PINTELAS, P. 2007. A Wrapper for Reweighting Training Instances for Handling Imbalanced Data Sets. In IFIP International Fede*ratio*n for Information Processing,Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C, Pnevmatikakis, L., Polymenakos, L., (Boston: Springer), pp. 29-36.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \quad (9)$$

Donde $R^2_{X_j|X_{-j}}$ corresponde al R^2 de una regresión de X_j sobre el resto de las variables explicativas[12]. Un valor mayor a 5 ó 10 implican un problema de multicolinealidad[13].

3.3.5 Regresión logística

La regresión logística es similar un modelo de regresión lineal pero adaptada a modelos donde la variable dependiente es dicotómica. Un coeficiente de regresión logística corresponde al logaritmo del *odds ratio* de cada variable independiente del modelo.

Para entender qué es un *odds ratio* se debe comprender primero lo que es un *odds*. El *odd* es la probabilidad de que suceda un evento dividido por la posibilidad que no suceda[14]. A modo de ejemplo, se tiene que la probabilidad de saturarse es 0,8. Luego, el *odd* de saturarse es 0,8/0,2=4.

Por su parte, el *odds ratio* asocia dos variables, en este caso, la variable dependiente con la independiente. Se busca establecer la fortaleza de la relación y si es positiva o negativa. Siguiendo con el ejemplo acerca de la saturación, donde la variable dependiente es si un cliente está saturado o no y se agrega una variable independiente que indica si el cliente está casado (1) o no (0), entonces el *odds ratio* se calcula usando los siguientes datos:

- Probabilidad de saturarse y estar casado es 0,75.
- Probabilidad de no saturarse y estar casado es 0,25.

Así, el *odd* de saturarse y estar casado es 0,75/0,25=3.

- Probabilidad de saturarse y no estar casado es 0,58.

[12] JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. 2017. An Introduction to Statistical Learning (8th ed.). Springer Science+Business Media New York. pp. 101-102.
[13] SHEATHER, S. 2009. A modern approach to regression with R. New York, NY: Springer.
[14] CÁRDENAS, J. 2015. *Odd ratio*: qué es y cómo se interpreta [En línea] < http://networkianos.com/*odd-ratio*-que-es-como-se-interpreta/#toc-1> [Consulta: 10 Enero 2019]

- Probabilidad de no saturarse y no estar casado es 0,42.

Así, el *odd* de saturarse y no estar casado es 0,58/0,42=1,38.

Por lo tanto, el *odd ratio* de saturarse y estar casado es 3/1,38=2,17.

Este número se interpreta como que cuando un cliente está casado los *odds* son 2,17 veces más grandes de saturarse que cuando no está casado. Cuando el *odd ratio* es 1, indica ausencia de relación entre las variables. Si es menor que 1, existe una asociación negativa entre las variables y cuando es mayor que 1 hay una asociación positiva.

Un modelo estimado con una regresión logística no es un clasificador por sí mismo, aunque se puede usar como uno al escoger un valor de corte para la variable dependiente. De esta forma, un valor por sobre el corte pertenece a una clase y otro por debajo a otra.

3.3.6 Métrica de desempeño

Dado que se busca calcular la probabilidad de saturación, se necesita un indicador que diga qué tan buen predictor es el modelo definido. Para evaluar esta capacidad predictiva del modelo se utiliza la curva ROC.

i. Curva ROC

La curva ROC, acrónimo de *Receiver Operating Characteristic*, es una medida de la performance de un modelo de clasificación en función del valor de corte definido para establecer una clasificación como de una clase u otra. Lo que entrega es qué tan capaz es el modelo de distinguir entre clases.

Para entender lo que sigue es necesario introducir los conceptos de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN):

- TP: ocurre cuando el modelo indica que la variable dependiente es 1 (positiva) y eso concuerda con la realidad.
- TN: ocurre cuando el modelo indica que la variable dependiente es 0 (negativa)

y eso concuerda con la realidad.

- FP: ocurre cuando el modelo indica que la variable dependiente es 1 (positiva) y eso no concuerda con la realidad.

- FN: ocurre cuando el modelo indica que la variable dependiente es 0 (negativa) y eso concuerda con la realidad.

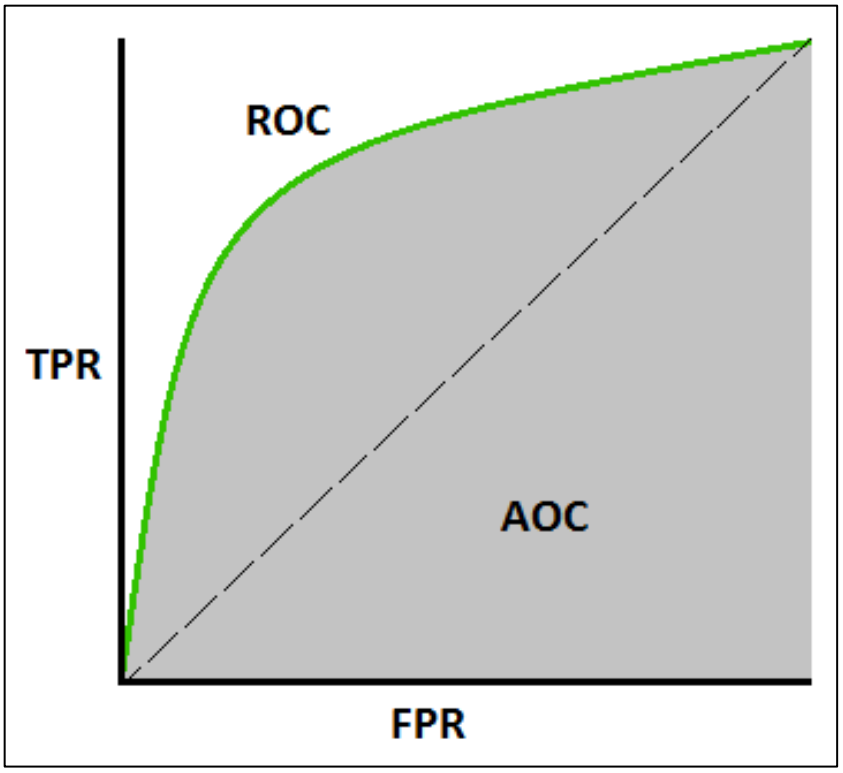

Ilustración 7. Curva ROC y sus componentes.
Fuente: *Towards Data Science*.

La ilustración muestra la curva ROC. Allí hay conceptos que se explican a continuación:

- TPR (*True positive rate*): también conocido como *recall* o sensibilidad, corresponde a la capacidad del modelo para dar como positivos los casos que realmente lo son. La fórmula es la siguiente:

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

- Especificidad: corresponde a la capacidad del modelo para dar como negativos

los casos que realmente lo son. La fórmula es la siguiente:

$$Especificidad = \frac{TN}{TN + FP} \tag{11}$$

- FPR (*False positive rate*): corresponde a una tasa que indica la probabilidad de que el modelo dé como positivo un caso que no lo es. Es el complemento de la especificidad. La fórmula es la siguiente:

$$FPR = 1 - Especificidad$$

$$= \frac{FP}{TN + FP} \tag{12}$$

La curva ROC representa la relación entre la TPR con la FPR frente a diferentes umbrales de clasificación. Con esto, se tienen muchos puntos en esta curva que se pueden evaluar con una regresión logística variando el umbral de clasificación, lo que es ineficiente. Por ello, se utiliza el *Area Under Curve* (AUC), que es una métrica agregada del rendimiento del modelo en todos los umbrales de clasificación posibles [15].

El AUC corresponde al área bajo la curva ROC. Si su valor es 1, entonces el modelo distingue perfectamente entre los casos positivos y los negativos. Si su valor es 0,5 se tiene un modelo sin capacidad de discriminar entre los casos positivos y negativos. Si su valor es 0 entonces el modelo predice los positivos como negativos y viceversa.

[15] Aprendizaje automático. Clasificación AUC y ROC. 2018. [18-01-2019]

IV. METODOLOGÍA

La metodología propuesta se basa en el proceso KDD (*Knowledge Discovery in Databases*) (Fayyad, 1996). Corresponde a un flujo de etapas diseñado para la extracción de conocimiento útil de grandes bases de datos. Cuenta con cinco etapas: selección, preprocesamiento, transformación, *data mining* y finalmente evaluación e interpretación. La ilustración 8 resume esta metodología.

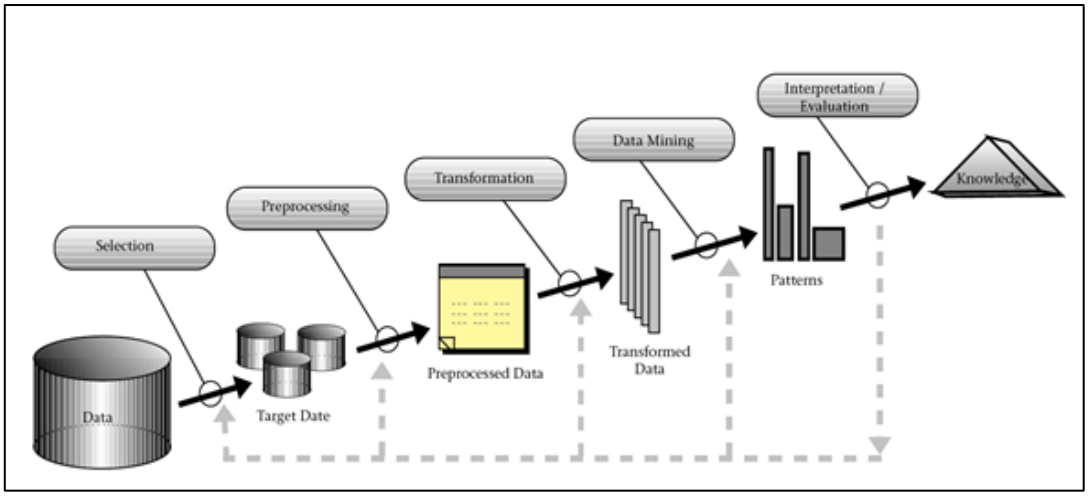

Ilustración 8. Flujo del proceso KDD.
Fuente: *Research Gate*.

4.1 Selección de datos

Esta etapa incluye el levantamiento de las distintas fuentes de información y los datos a utilizar durante el trabajo. En este caso, dado que la memoria se enmarca en el *email marketing*, se trabaja con tablas de envíos de *email* por cliente, de envío de *email*s masivos y de remitentes. También se incluye información sociodemográfica y transaccional de los clientes. Cabe destacar que no sólo se usa información del negocio donde se estudia la saturación, sino que también de otros negocios del *holding*.

4.2 Preprocesamiento

En primer lugar, se realiza un análisis exploratorio de los datos y un estudio de informes elaborados por personas encargadas del *email marketing* a nivel de *holding* para desarrollar un entendimiento del tema. Se hace uso de estadísticos descriptivos

como medias, medianas y gráficos. En el proceso, es posible encontrarse con campos de datos vacíos o erróneos. Un registro con estas características se debe descartar o corregir si es posible, para así no ensuciar el modelamiento.

En segundo lugar, se realiza un análisis de *outliers* para identificar y descartar observaciones atípicas que puedan ensuciar la calibración de los modelos. Se usa el *test* de Tukey y criterios de negocio para esta parte. El *test* nombrado usa el concepto de rango intercuartílico, que es la diferencia entre el tercer y el primer cuartil. Cuando un dato se encuentra bajo el primer cuartil menos 1,5 veces el rango intercuartílico o sobre el tercer cuartil más 1,5 veces el rango intercuartílico, se considera *outlier* en principio.

4.3 Transformación

Luego de pre-procesar los datos, se debe realizar la transformación de variables. Esto corresponde a la creación de nuevas variables que resuman la información entregada desde los datos crudos, con el objetivo de poder trabajar las variables y facilitar el objetivo del trabajo. Por ejemplo: agregar la cantidad de *emails* recibidos por cliente a nivel diario, transformar variables en formato *string* en un formato numérico, definir variables binarias, etc.

Para apoyar el proceso creativo de la generación de variables, es de gran utilidad plantear hipótesis sobre qué variables pueden ser relevantes a la hora de entender la saturación. Por ejemplo, el tipo de cliente en algún negocio es inicialmente una variable categórica que se debe transformar a tantas variables binarias menos uno, que corresponde a la línea base. Al momento de segmentar, se normaliza cada una de las variables de interés para estandarizar las escalas.

4.4 *Data mining*

Luego de pre-procesar y trabajar los datos, se debe desarrollar el modelamiento. A continuación, se explica cómo se segmenta a los clientes del negocio en función de las variables de interés para el *email marketing*, donde se considera a la cantidad de envíos y la tasa de apertura. Luego, el proceso para escoger un criterio de saturación mediante la generación de distintos escenarios. Posteriormente, la generación de un modelo probabilístico que permita calcular la probabilidad de saturación y finalmente la propuesta de un costo de envío de *emails* que permita al *holding* en un futuro implementar un sistema que incentive el buen uso de este canal de *marketing*.

4.4.1 Segmentación de clientes

El negocio no envía *emails* de forma aleatoria, sino que considera, tal como lo hace la actual política de toques, el comportamiento de lectura. Por tanto, dado que se quiere conocer a los clientes en función de su comportamiento ante el *email marketing*, no basta con conocer si un cliente abre los correos, sino que también la cantidad de *emails* que recibe. Por lo tanto, las variables usadas son la cantidad de envíos y la tasa de apertura de envíos de *emails* enviados por el negocio en estudio a sus clientes durante el segundo semestre del año 2017.

Se descarta segmentar por variables transaccionales ya que el objetivo de la segmentación es lograr posteriormente definir un criterio de saturación, lo cual depende del comportamiento de lectura de *emails* de las personas. Puede existir el caso de personas con distinto comportamiento transaccional y con un mismo comportamiento ante los *emails*. Posterior a este análisis se estudia el impacto transaccional.

Para segmentar a los clientes por las variables mencionadas, primero éstas se normalizan, dado que una de ellas corresponde a una probabilidad (va entre 0 y 1) y otra es una variable no negativa y discreta. Posteriormente, se usa el método de *k-medias* para obtener los segmentos. El criterio usado para definir la cantidad de segmentos es el *Elbow method* (Thorndike, 1953), donde se escoge el punto donde la variación en la inercia sea menor, lo que se observa gráficamente como un cambio brusco en la inercia.

4.4.2 Definición de saturación

Existe la posibilidad de que la baja en la tasa de apertura afecte a las personas que compran preferentemente a través de la plataforma *web*. Esto se ve apoyado por un estudio que muestra que el factor más relevante asociado a altas tasas de apertura es la compra *online*. Esto indica que existe una relación entre las altas tasas con las compras *online*. Se indica también que la expansión de la Internet va de la mano con el aumento en las tasas, pero que el mal uso del *email marketing* puede socavar su desarrollo (Chittenden, Rettie; 2003).

Dado lo anterior, se estudia a los clientes considerando solamente sus transacciones *web*. Además, se hace el ejercicio de considerar a clientes con distintos grados de adopción de las transacciones *web*, es decir, con distintos porcentajes de transacciones *web* sobre el total de sus compras. La hipótesis detrás de esto es que los clientes a partir de cierto grado de adopción *web* pueden mostrar

una relación entre la caída en sus tasas y sus transacciones *web*.

i. Definición del grado de adopción *web*

Se debe definir qué grado de adopción *web* exigir a los clientes a considerar en cada uno de los segmentos. La adopción *web* corresponde al porcentaje de transacciones realizadas a través del canal *web* del negocio sobre el total de las transacciones hechas en el mismo. Se usa un año de información para determinar el grado de adopción *web*, desde octubre de 2016 a septiembre de 2017, dado que desde octubre de 2017 se usa información para el modelo predictivo.

Para poder definir el grado de adopción *web* se estudia cómo varía el efecto de saturarse, a través de la diferencia de *lifts* de transacciones, para un escenario dado. Se utilizan fenómenos cuyas características son que corresponden a una baja considerable en la tasa (sobre un 70%) y por un período no despreciable. Con esto definido, el siguiente paso es escoger un fenómeno de saturación para cada segmento.

ii. Definición del criterio de saturación

Para definir el criterio de saturación se requiere de dos conceptos definidos en la justificación del proyecto. Uno es la caída en la tasa y la temporalidad. La caída en la tasa puede ser relativa o absoluta. Cuando se habla de una caída relativa se refiere a una caída porcentual respecto a la tasa de referencia. Por su parte, una caída absoluta corresponde a caer por debajo de un umbral independiente de la tasa de referencia, por ejemplo, llegar a una tasa de apertura de un 10%.

La temporalidad se estudia por semanas, puesto que no es factible tomar el día como referencia ya que es muy poco tiempo considerando que muchos clientes no reciben *emails* a diario. Tomar un período más largo tampoco es conveniente porque se puede perder la posibilidad de detectar el momento donde se genera la saturación. Por último y más relevante, se usa la semana porque la política de toques, que es la que rige los envíos, lo hace.

Un fenómeno corresponde a definir una caída en la tasa sostenida en un período de tiempo. Cuando lo anterior ocurra, entonces se considera al cliente como saturado. Cabe destacar que se busca un criterio de saturación para cada segmento de clientes, ya que tienen distintas características con respecto al *email marketing* y, por tanto, no son comparables.

Además de observar una caída en la tasa de apertura se busca que asociada a ésta exista un impacto en variables transaccionales del cliente. Esto se busca porque para la empresa es de gran utilidad conocer no sólo cuando sus clientes dejan de leer sus campañas, sino si además esto hace que un cliente gaste menos dinero o compre con menor frecuencia en el negocio.

Para entender si existe un impacto en variables transaccionales del cliente se generan distintos escenarios. Los escenarios se configuran a partir de sensibilizar la caída en la tasa y la duración de ésta (que en conjunto constituyen el fenómeno).

La variable considerada son las transacciones. Se desestima el uso del gasto ya que, por un lado, las transacciones están altamente correlacionadas con este indicador. Por otro lado, las transacciones son más representativas de la situación de una persona que se satura y, por tanto, tiene una apreciación negativa del negocio. Por ejemplo, una persona podría comprar un refrigerador, lo que aumentaría su gasto, pero si sólo compró eso en el negocio y el resto de sus productos en la competencia, se pierde el sentido buscado.

Dicho esto, se tiene la cantidad de transacciones de las personas a un, tres y seis meses antes y después de haber tenido algún fenómeno. En cada escenario se comparan aquellas personas a las que les ocurre el fenómeno respecto a las que no. Esto se realiza para los meses de marzo, abril y mayo. Se utiliza 1 mes dado que se busca atribuir la caída a la saturación y si se considera una ventana temporal mucho más amplia se puede decir que la caída puede deberse a otros motivos.

La idea de generar estos escenarios es entender cuándo hay un impacto en las transacciones, para luego confirmarlo realizando una regresión que permita saber si hay un efecto estadísticamente significativo.

Se espera elegir un fenómeno que incida sobre la diferencia de transacciones en un mes atribuido a saturarse y que cuente con el criterio menos estricto (combinación de cantidad de semanas y caída en la tasa). Se usa un mes ya que se observa que se encuentra altamente relacionado con lo que ocurre a tres y seis meses y además muestra los impactos de mayor magnitud.

Para comparar los resultados de los “saturados” versus los “no saturados” en cada escenario se usa la herramienta de *diferencias en diferencias* para conocer el efecto de la saturación. Específicamente, se calcula el *lift* de transacciones para los “saturados” y los “no saturados” del mes anterior al mes posterior, es decir, qué porcentaje es la diferencia de las transacciones con respecto a las transacciones del mes anterior. A continuación, se expresa con una fórmula:

$$\% \, Lift = \frac{Trx_{post} - Trx_{pre}}{Trx_{pre}} * 100 \quad (13)$$

Posteriormente, se calcula la diferencia de *lifts* entre el grupo saturado y el no saturado. De esta manera, cuando la diferencia es negativa, entonces hubo un impacto en las transacciones por el hecho de haberse saturado, ya que se está controlando la estacionalidad al calcular la diferencia de las diferencias.

Para evaluar qué escenarios generan potencialmente más impacto se utiliza un indicador. Este indicador corresponde al producto entre la diferencia de diferencias de transacciones (efecto del tratamiento) con la cantidad de personas que les ocurre el fenómeno, que llamaremos indicador de impacto.

Luego, se verifica a través de regresiones, en su rol descriptivo, la significancia estadística de los fenómenos que tengan un menor indicador de impacto (los más negativos). Esto, porque sólo se estudia comparativamente el efecto de saturarse en los diferentes escenarios, sin entender si realmente existe un efecto.

La regresión usada para estudiar el efecto de los distintos fenómenos es la siguiente:

$$\begin{aligned} \triangle \boldsymbol{Trx´s\ 1\ mes\ retail} &= \alpha + \beta_{fen\_duración} * dummy\ fenómeno \\ &+ \sum \beta_i * vars\ transac\ retail\ pre_i \\ &+ \sum \beta_j * vars\ transac\ \ tarjeta\ pre_j \\ &+ \sum \beta_k * vars\ nav\ web\ pre_k \\ &+ \sum \beta_l * vars\ sociodemográficas_l + \varepsilon \end{aligned} \quad (14)$$

4.4.3 Desarrollo de modelo predictivo

Se busca desarrollar un modelo que pueda predecir la probabilidad de saturación de un cliente para un día dado. Para esto, se cuenta con seis grupos de variables, sobre las cuales se profundiza en el desarrollo metodológico. Dado que a priori se

desconoce la relevancia de estas variables para predecir la saturación, se realiza una selección de variables. Además, se genera la variable dependiente binaria que corresponde a si el cliente se satura o no un día dado. Ésta tomar el valor 1 si se satura y 0 si no.

La selección de variables consiste en aplicar un método de mínimos cuadrados penalizados llamado SCAD, el cual indica qué variable se puede descartar del modelo pues no aporta al ajuste de los datos. Posteriormente, con las variables restantes se prueba el nivel de ajuste del modelo a los datos usando como métrica el AIC. Se prueba sacando las variables menos significativas y observando si mejora el AIC. Esto corresponde a un método *backward*. Cuando se llega a un momento en que seguir sacando variables no mejora este indicador, sino que lo empeora, se detiene el proceso y se elige el modelo con el menor AIC.

Posteriormente, se estudian posibles problemas de multicolinealidad, dada la similitud entre algunas variables. Para esto se usa el VIF. Luego, se balancea la base de entrenamiento buscando tener una calibración que permita optimizar la predicción del modelo[16]. Finalmente, se prueba la capacidad predictiva del modelo usando como métrica el AUC, que corresponde al área bajo la curva ROC.

4.4.4 Propuesta de costo de envío

Se propone el cálculo de un costo de envío. El objetivo de este costo es incentivar a los negocios dentro del *holding* a no abusar del uso del *email marketing*, dado que se observa que se puede provocar la saturación de un cliente y en consecuencia un impacto transaccional o bien otro efecto no tan directo.

Los aspectos que influyen en el costo de envío de un *email* son la probabilidad de saturación y el potencial impacto transaccional sobre el cliente. Dado que este cálculo es una propuesta inicial y depende del cálculo del resto de los negocios, tan sólo se especifica cómo se relacionan los aspectos influyentes.

A través del modelo de saturación se tienen las variables que permiten calcular la probabilidad de saturación que hacen variar el costo de envío dependiendo de su valor. Por otro lado, el impacto transaccional puede ser mayor o menor en la medida del grado de adopción *web* del cliente. Por lo tanto, a mayor grado de adopción *web*, mayor costo por saturar al cliente y por tanto mayor costo de envío.

[16] DROZDOWICZ, B. EVIN, D y HADAD, A. 2007. Modelo para el Tratamiento de Datos Desbalanceados basado en Redes Neuronales Autoorganizadas. Universidad Nacional de Entre Ríos, Facultad de Ingeniería. pp 1-3

V. ALCANCES Y RESULTADOS ESPERADOS

5.1 Alcances

- Estudio de carácter observacional, no considera experimentos ni encuestas para obtener datos. El trabajo se basa en datos disponibles de la empresa de los años 2017 y 2018.

- Formalización del concepto de saturación en el negocio de *retail* en estudio.

- No se consideran *emails* de carácter operacional o de prueba, dado que no se incluyen en una planificación de toques.

- Fuera del alcance se encuentra el desarrollo de un sistema de costo en el envío de *emails* para todos los negocios del *holding*. El objetivo de este sistema es generar incentivos al buen uso del *email marketing*, evitando la saturación de los clientes.

5.2 Resultados esperados

- Cálculo de una probabilidad de saturación diaria para los *emails* del negocio de *retail* en estudio para cada segmento de clientes definidos que permita tomar decisiones dinámicas respecto al envío de *emails*. Actualmente, existe una política de toques fija.

- Propuesta de una fórmula de costo de envío a partir de la probabilidad de saturación diaria e impacto transaccional para el negocio de *retail* en estudio. El objetivo de asignar costos al envío de *email*s es incentivar el buen uso de la herramienta dentro del *holding*. Dado que se requiere conocer qué ocurre con el resto de los negocios, sólo se identifican los factores que influyen sobre el cálculo de este costo.

- Generación de *insights* para mejorar las prácticas en *email marketing*.

VI. DESARROLLO METODOLÓGICO

6.1 Preparación de los datos

6.1.1 Selección de los datos

El *holding* cuenta con un código interno para identificar a sus clientes. Dicho esto, se considera a aquellos *emails* cuyo identificador esté asociado a sólo un código interno. Esto se debe a que se busca tener la certeza de que un correo es utilizado por un único cliente, ya que al momento de agregar variables sociodemográficas es necesario tener una relación unívoca y no generar registros duplicados para un mismo correo. Se tiene entonces una cantidad de 8.083.848 correos considerados inicialmente asociados a un código interno único, que no es lo mismo que el rut.

La información de envío de correos utilizada corresponde a los años 2017 y 2018. Con esto se cuenta con información suficiente para rescatar estacionalidades. Las fuentes de información usadas son las siguientes:

- Tabla de envíos: cada registro corresponde al envío de un *email* a un cliente en particular. No se considera un *email* de tipo *trigger*[17] dentro de los *emails* del negocio en estudio, puesto que, dado sus características, tiene en general una probabilidad de apertura mayor que los correos de campañas.

- Tabla de envíos masivos: cada registro corresponde a un mismo *email* que se envía a varios clientes. Esto puede ser el caso de una campaña.

- Tabla de remitentes: esta tabla relaciona el código del remitente del *email* que aparece en la tabla de envíos con el nombre del negocio remitente. No se considera un remitente que corresponda a correos de prueba u operacionales. Un ejemplo de estos últimos es cuando se avisa la compra de un producto a un cliente. Esto, porque un *email* de esas características no forma parte de la política de toques.

- Tablas con información transaccional: se cuenta con tablas de los distintos negocios con información transaccional. Además, hay tablas con las cuales se puede filtrar la información útil, como por ejemplo considerar sólo ventas y no devoluciones o diferenciar ventas físicas de ventas *web*.

- Tablas con información de navegación *web*: existen tablas con información de navegación en la página *web* del negocio de *retail*. Se cuenta con

[17] Corresponde a un *email* que se envía basado en un evento, como un cupón para el día de cumpleaños de un cliente.

variables tales como si ve, agrega a la bolsa u ordena algún producto.

- Tabla con información sociodemográfica: se cuenta con datos tales como la edad, el GSE, sexo, estado civil, entre otras.

La ilustración siguiente muestra las fuentes de información indicando, en el caso de las tablas de *email marketing*, los distintos tipos de información almacenados y que son transversales a los negocios. En el caso de las transaccionales y de navegación *web* se indican los negocios incluidos. Por su parte, las sociodemográficas son transversales, ya que son relativas a los clientes del *holding* y no a un negocio en particular.

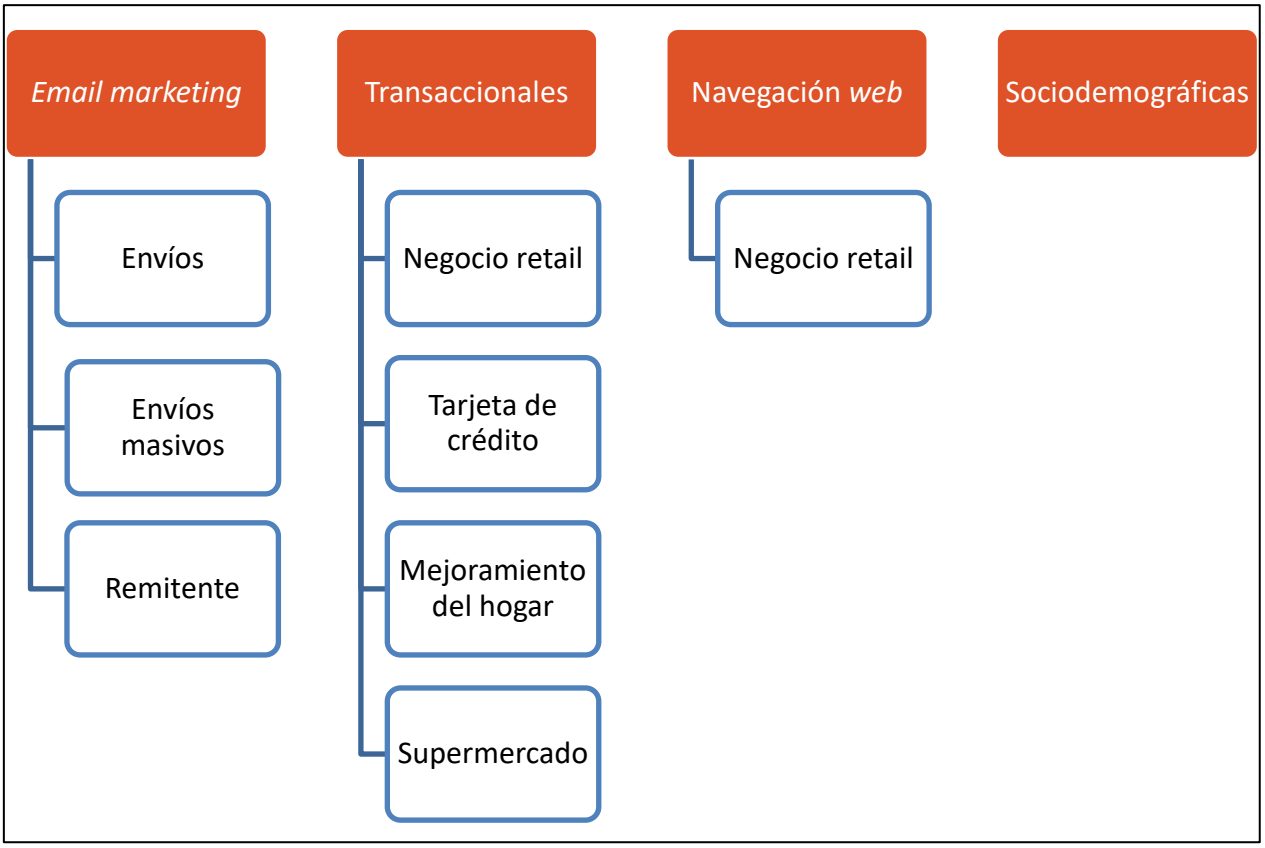

Ilustración 9. Fuentes de información.
Fuente: Elaboración propia.

6.1.2 Valores faltantes o erróneos

En la variable rut se tiene que hay valores incorrectos o que no hay que considerar, ya sea porque se hizo una mala ingesta de esos datos o porque corresponde al rut de una empresa.

No se consideran aquellas variables que cuenten con más de un 30% de valores faltantes. Ellas corresponden a las variables transaccionales del negocio de mejoramiento del hogar y del negocio de supermercados.

Para el resto de las variables, en caso de tener un valor faltante se imputa la mediana de la variable correspondiente. El caso más significativo corresponde a las variables transaccionales del negocio de *retail*, las cuales cuentan con un 25% de valores faltantes a los cuales se les imputa la mediana. En la siguiente tabla se muestran los tipos de variables junto al porcentaje de valores faltantes:

Tipo de variable	Valores faltantes
Transaccional *web* mejoramiento del hogar	83%
Transaccional supermercado	68%
Transaccional mejoramiento del hogar	38%
Transaccional negocio *retail*	25%
Email supermercado	21%
Email mejoramiento del hogar	19%
Transaccional *web* tarjeta crédito	15%
Email tarjeta de crédito	7%
Navegación *web* negocio *retail*	5%
Transaccional *web* negocio *retail*	3%
Transaccional tarjeta crédito	2%
Email negocio *retail*	2%
Sociodemográficas	0%

Tabla 5. Valores faltantes para distintos tipos de variables.
Fuente: Elaboración propia.

También se tiene que ciertas variables transaccionales, como el gasto, tienen observaciones negativas. Esto puede ocurrir por devoluciones y representa un 1,95% de los registros, los cuales no se consideran.

6.1.3 *Outliers*

En primer lugar, se analizan las variables usadas para segmentar. Existe la posibilidad de considerar personas que hayan recibido muy pocos *emails* durante el período de segmentación, que corresponde al segundo semestre de 2017. A modo de ejemplo, se tiene un caso de un correo que recibe sólo un *email* durante el período de segmentación y fue abierto.

Esto puede llevar a problemas en los segmentos, dado que se tiene a una persona que abre todo lo que recibe, es decir, tiene una tasa de apertura de 100%, pero que no debe ser considerada si no recibe *emails* de manera frecuente. La cantidad de clientes que recibe al menos un *email* durante el período de segmentación son 5,7 millones, es decir, que recibieron por lo menos un *email* en un período de seis meses.

Por lo tanto, se debe imponer un requisito a los correos a considerar en la segmentación. Éste no puede ser muy exigente, para así tener una cantidad suficiente de clientes y lograr una segmentación representativa de la población. Con esto en mente y usando criterios de negocio, se tiene la opción de considerar correos que hayan recibido cierta cantidad de *emails* del negocio cada uno de los meses del período de segmentación.

La siguiente ilustración muestra la cantidad de clientes que se consideran si el criterio de ingreso a la segmentación varía, yendo desde un *email* hasta seis cada mes:

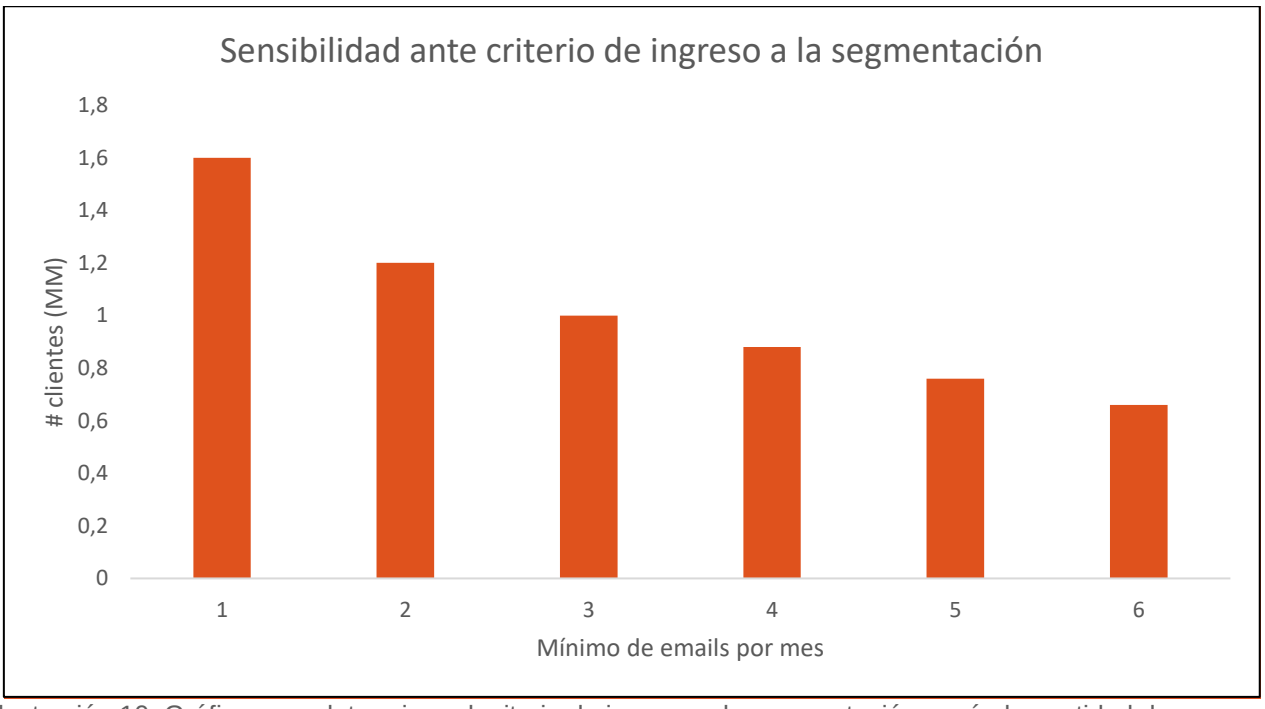

Ilustración 10. Gráfico para determinar el criterio de ingreso a la segmentación según la cantidad de clientes considerados.
Fuente: Elaboración propia.

Con esto, se decide considerar a los correos que hayan recibido por lo menos un *email* durante cada uno de los meses del período de segmentación. Por tanto, se consideran 1.601.350 clientes.

Además, se utiliza el *test* de Tukey para eliminar *outliers* a través de un análisis univariado. Se encuentran clientes cuyas variables tales como el gasto o las transacciones son muy altas con respecto al resto. Estos clientes corresponden a un 1,82% y se descartan. Por lo tanto, quedan 1.572.205 clientes.

6.1.4 Transformación de variables

Se realiza una agrupación de los registros de la tabla de envío de *emails* a nivel diario, donde se tiene información acerca de recepciones, aperturas y clicks de *emails*. El resultado de esto es tener, para un cliente y día dado, la cantidad de correos recibidos, abiertos y clickeados.

Se transforma también la variable “asunto”, que contiene un *string* con el contenido del asunto del *email*, a una variable llamada “largo del asunto”, que contiene la extensión del asunto. Se generan también las variables “pc” y “mobile” a partir de

la variable “sistema operativo”. Estas son variables *dummies* que indican si el dispositivo desde el cual se recibe el *email* es un computador o un teléfono móvil.

Se generan variables *dummies* a partir de otras como el género, donde en caso de ser género masculino se tiene un 1 y 0 en caso de ser femenino. Lo mismo ocurre con el estado civil, donde estar casado se refleja con un 1 y un 0 en caso de estar soltero. Con las regiones se recoge el hecho de pertenecer a la región Metropolitana (1) considerando como base pertenecer a otra región (0).

Finalmente, se generan las variables relacionales, las cuales permiten considerar el hecho de que la probabilidad de apertura de un mail en un momento dado está influida por los eventos que ocurrieron anteriormente. Por ejemplo, variables tales como la cantidad de *emails* que un negocio determinado envía el día anterior o los últimos siete días.

El total de variables con las que se trabaja finalmente son 255. La tabla siguiente muestra algunos ejemplos de variables agrupadas por tema. El total de las variables se encuentra en el anexo 1.

Mail	Sociodemográficas
Largo del asunto	Soltero
Pc*	Casado
Teléfono móvil*	Masculino
Temporales	Femenino
Mes	Edad
Semana	**Relacionales**
Día	# *email*s recibidos 5 días previos de negocio det.
Transaccionales	# *emails* abiertos 5 días previos de negocio det.
Gasto en 1 mes previo en un negocio det.	# *emails* clickeados 5 días previos de negocio det.
Visitas en 1 mes previo en un negocio det.	**Navegación *web***
Transacciones en 1 mes previo en un negocio det.	Cantidad de vistas en página 1 mes previo

Tabla 6. Ejemplos de variables generadas.
Fuente: Elaboración propia.

6.2 Segmentación de clientes

De manera de considerar la heterogeneidad de los clientes en términos de su comportamiento ante el *email marketing*, se realiza una segmentación en base a la cantidad de *emails* que reciben y su tasa de apertura, que es el criterio definido en el área corporativa. La información usada corresponde al negocio de *retail* en estudio durante el segundo semestre de 2017.

Se utiliza la técnica de *k-medias* para la segmentación. En la ilustración siguiente se observa cómo varía la inercia ante la consideración de distintos números de *cluster*s.

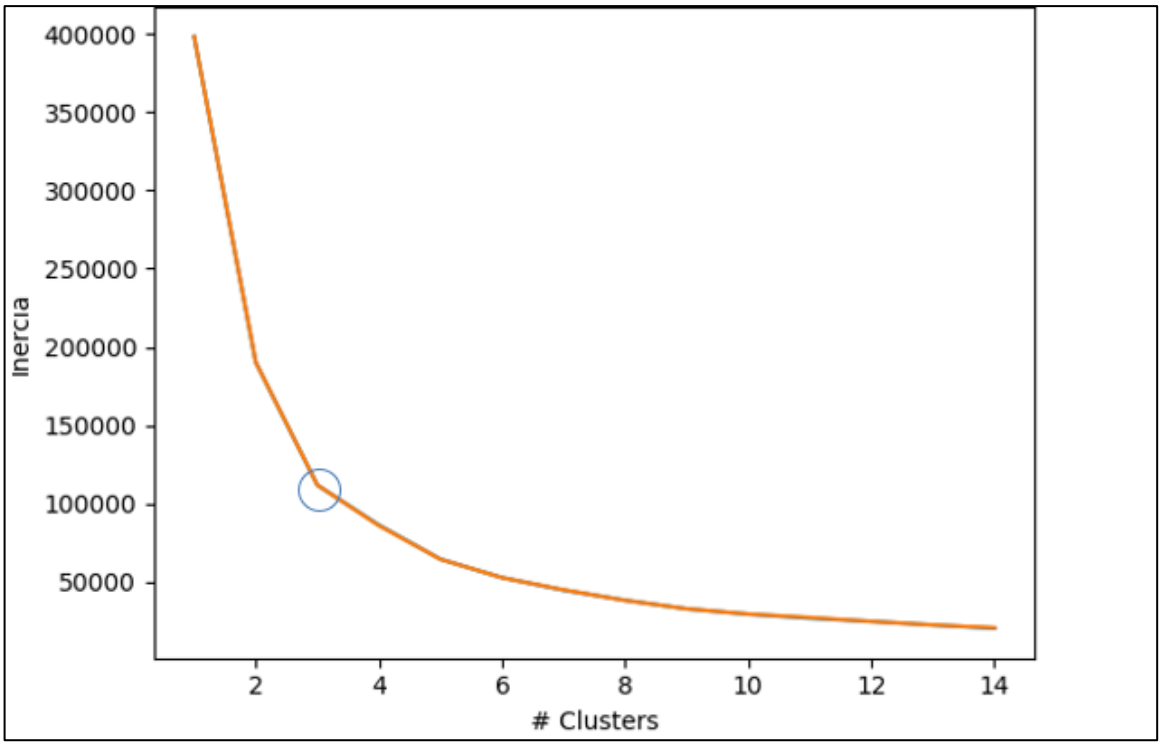

Ilustración 11. Variación de la inercia con respecto al número de *cluster*s.
Fuente: Elaboración propia.

Considerando el uso del *elbow method* y criterios de negocios, tales como la representación simple de los grupos, se consideran tres segmentos. En la siguiente ilustración se representan los segmentos con respecto a los centroides. El tamaño de la circunferencia representa la cantidad de personas que conforman el segmento.

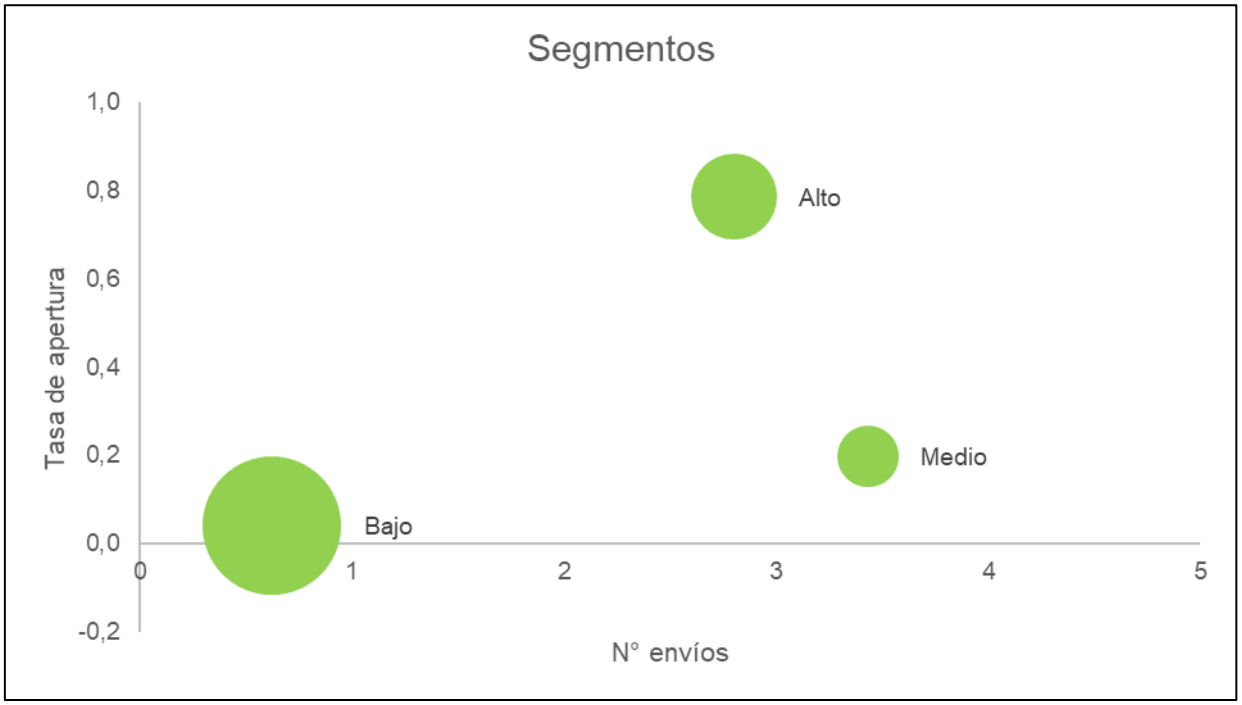

Ilustración 12. Representación gráfica de los segmentos.
Fuente: Elaboración propia.

La tabla siguiente resume algunas de las características de los segmentos. Nótese que los clientes se clasifican en alto, medio y bajo según su tasa de apertura para luego identificarlos más fácilmente.

Segmento	Centroide-envíos semanales	Centroide-tasa de apertura	% del total	# personas
Bajo	0,62	4%	63,29%	817.462
Medio	3,42	20%	24,51%	466.688
Alto	2,80	79%	12,20%	292.055

Tabla 7. Estadísticas de los segmentos.
Fuente: Elaboración propia.

A continuación, se muestran gráficos que representan el promedio de la tasa de apertura y el promedio de la cantidad de envíos recibidos semanalmente durante el primer semestre de 2018 para los distintos segmentos.

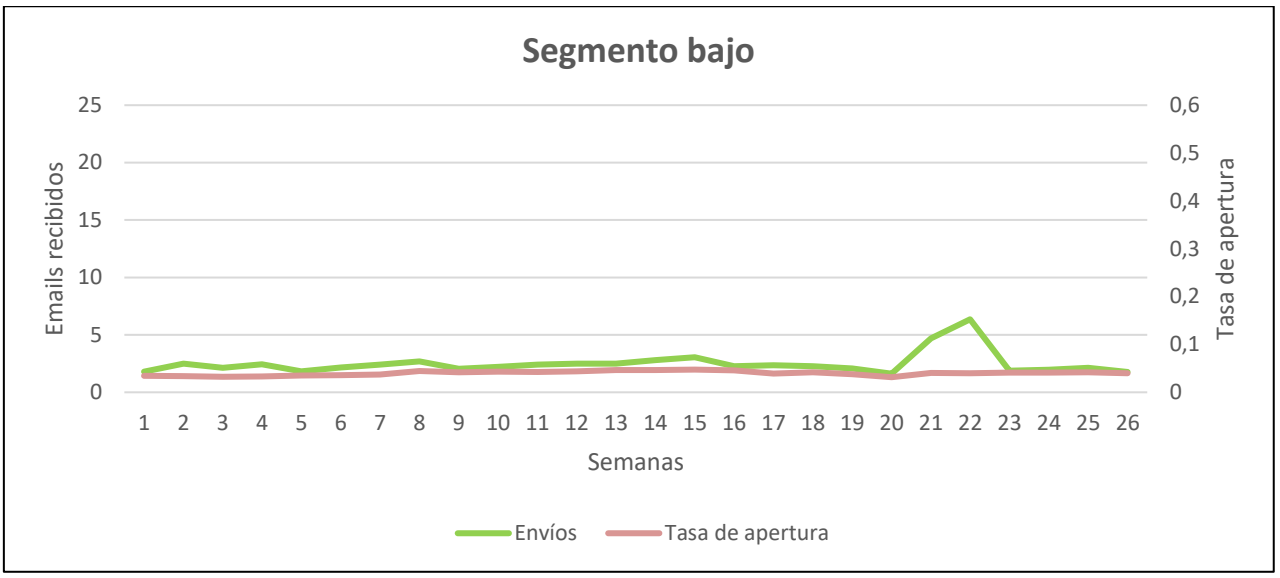

Ilustración 13. Evolución semanal primer semestre 2018 del segmento bajo.
Fuente: Elaboración propia.

Se observa que el segmento bajo presenta tasas de apertura muy cercanas a cero y pocos envíos (dos en promedio). Existe un *peak* de envíos la semana del 28 de mayo de 2018, que corresponde a la semana 22 de ese año.

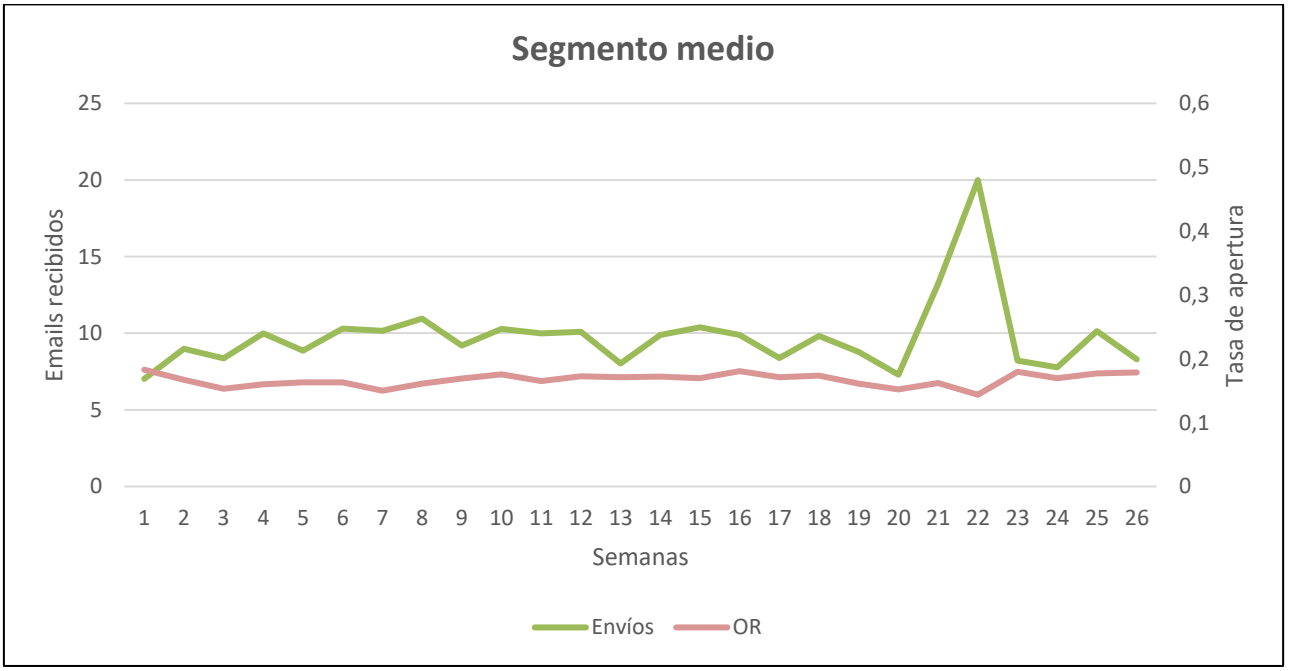

Ilustración 14. Evolución semanal primer semestre 2018 del segmento medio.
Fuente: Elaboración propia.

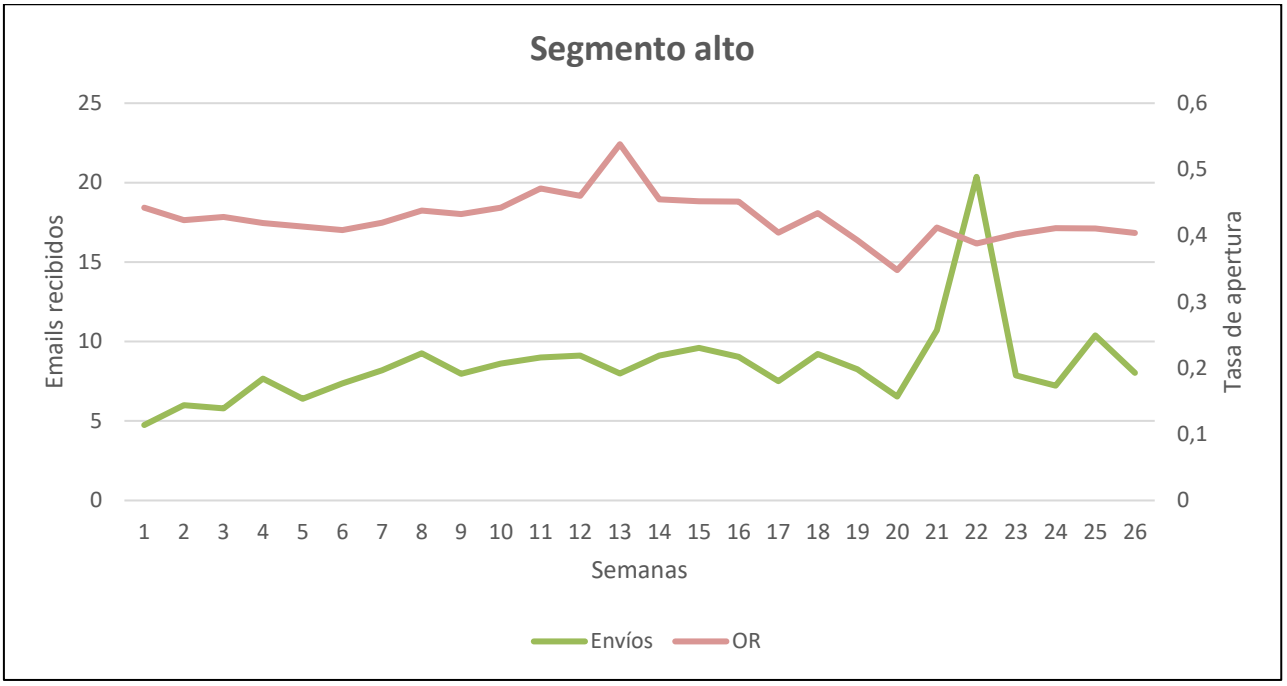

Ilustración 15. Evolución semanal primer semestre 2018 del segmento alto.
Fuente: Elaboración propia.

Se observan tasas de apertura entre 40% a 50% y envíos que casi alcanzan los 10 semanales en el segmento alto. Por otro lado, el segmento medio tiene tasas de apertura del orden del 16% y cantidad similar de envíos. Ocurre el mismo *peak* que en el segmento bajo por el mismo motivo.

6.3 Definición de saturación

Se generan escenarios donde se sensibilizan la caída en la tasa y la temporalidad. Para el segmento bajo, dado que sus tasas de apertura son cercanas a 0, se considera como caída en la tasa el dejar de leer *emails*, es decir, tener una tasa de apertura 0. Para los segmentos medio y alto se utilizan tasas relativas en la caída.

Se observa que los distintos escenarios observados para los segmentos no muestran caídas significativas debido a la saturación (presentan *lifts* cercanos a 0) ni una relación entre los resultados que evidencie una relación entre saturarse y presentar una baja en las transacciones, lo que en resumen se entiende como que no hay saturación que implique un impacto transaccional, que es lo que se busca. En general se descarta abril ya que, en mayo hay un *cyber* y esto altera los resultados. En anexos B se pueden observar estos los resultados.

Lo anterior implica que no hay una relación entre la saturación y una caída

transaccional cuando se consideran todos los canales para comprar. Sin embargo, se advierte una relación entre las tasas de apertura y las compras *online* (Chittenden, Rettie; 2003).

Dado esto, se tiene la hipótesis de que puede existir una relación entre saturarse y disminuir las transacciones si los clientes son preferentemente *web*, es decir, cuando sus transacciones son predominantemente a través de este canal. Se comienza el estudio con el segmento alto.

6.3.1 Segmento alto

i. Definición del grado de adopción *web*

Se estudia entonces cómo varía el efecto de saturarse, a través de la diferencia de *lifts* de transacciones, para un escenario dado variando el grado de adopción de este canal. Se utilizan fenómenos cuyas características son que corresponden a una baja considerable en la tasa (sobre un 70%) y por un período de al menos 3 semanas, que es donde se observan más efectos sobre las transacciones. Las ilustraciones 15,16 y 17 muestran lo descrito.

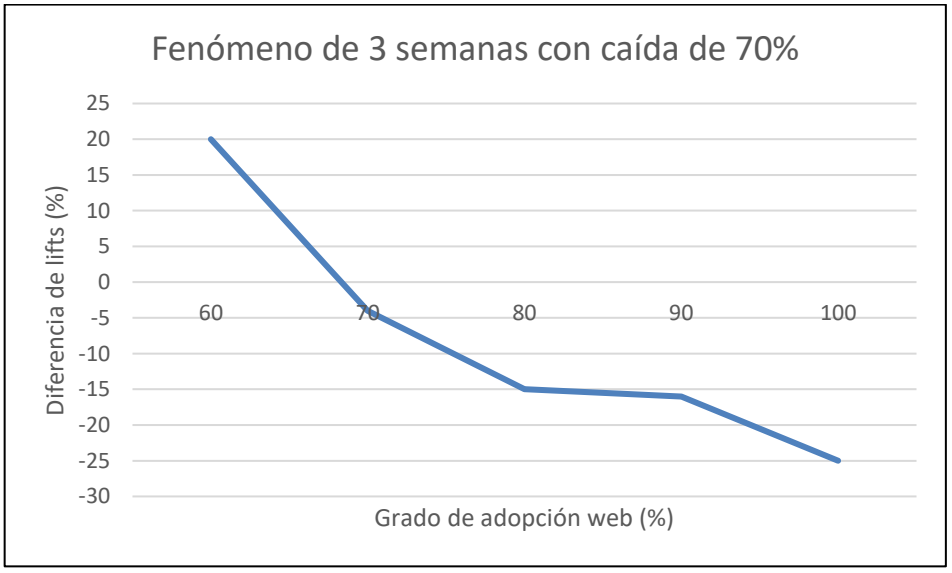

Ilustración 16. Definición del grado de adopción *web* para el fenómeno de 3 semanas y caída del 70% en la tasa.
Fuente: Elaboración propia.

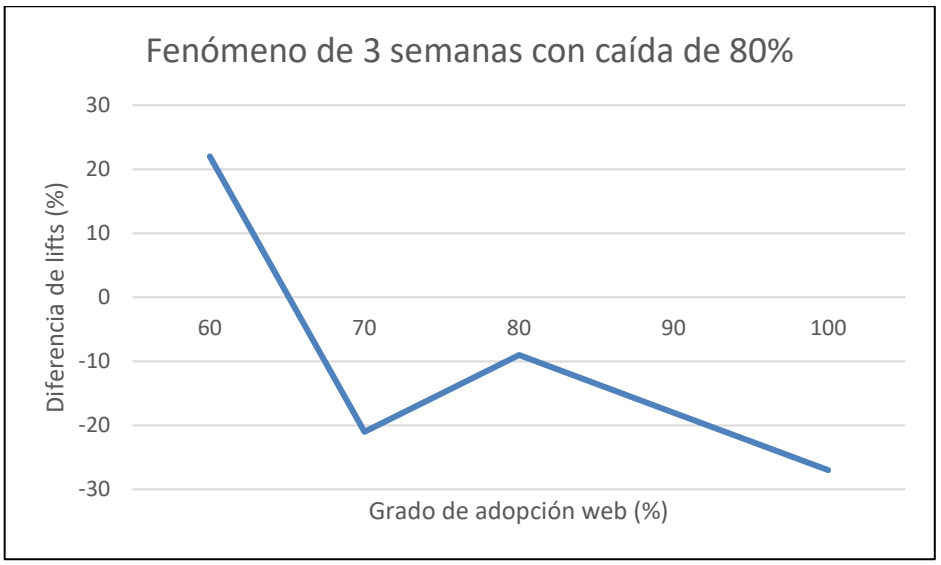

Ilustración 17. Definición del grado de adopción *web* para el fenómeno de 3 semanas y caída del 80% en la tasa.
Fuente: Elaboración propia.

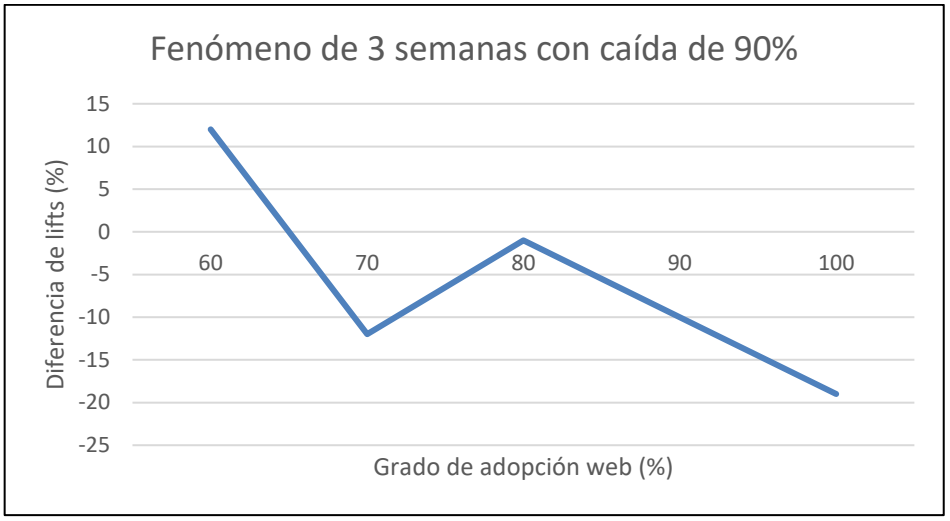

Ilustración 18. Definición del grado de adopción *web* para el fenómeno de 3 semanas y caída del 90% en la tasa.
Fuente: Elaboración propia.

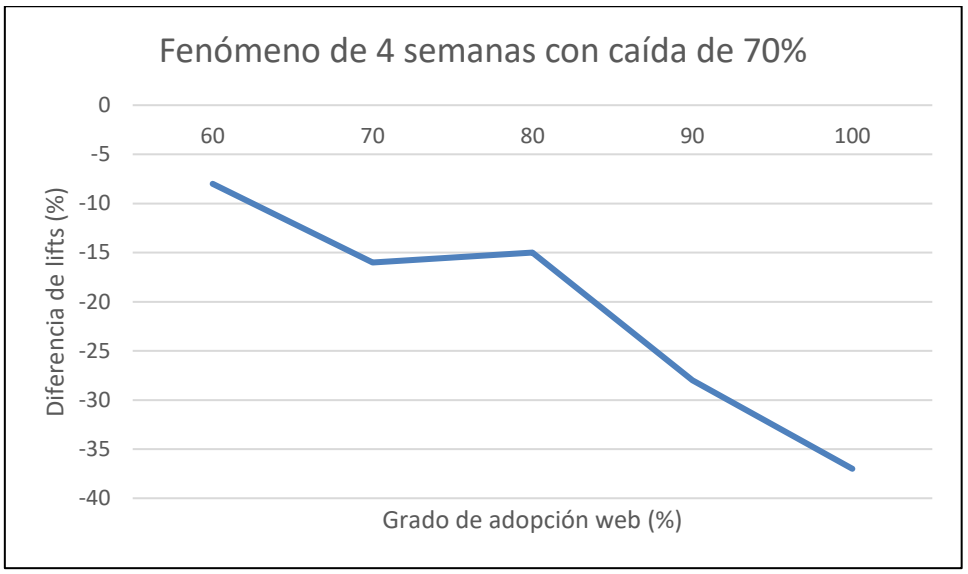

Ilustración 19. Definición del grado de adopción *web* para el fenómeno de 4 semanas y caída del 70% en la tasa.
Fuente: Elaboración propia.

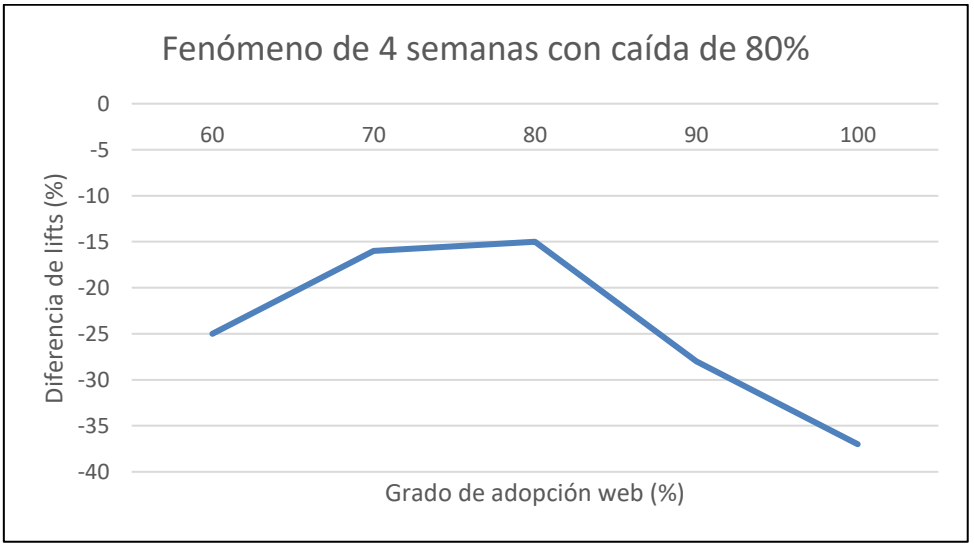

Ilustración 20. Definición del grado de adopción *web* para el fenómeno de 4 semanas y caída del 80% en la tasa.
Fuente: Elaboración propia.

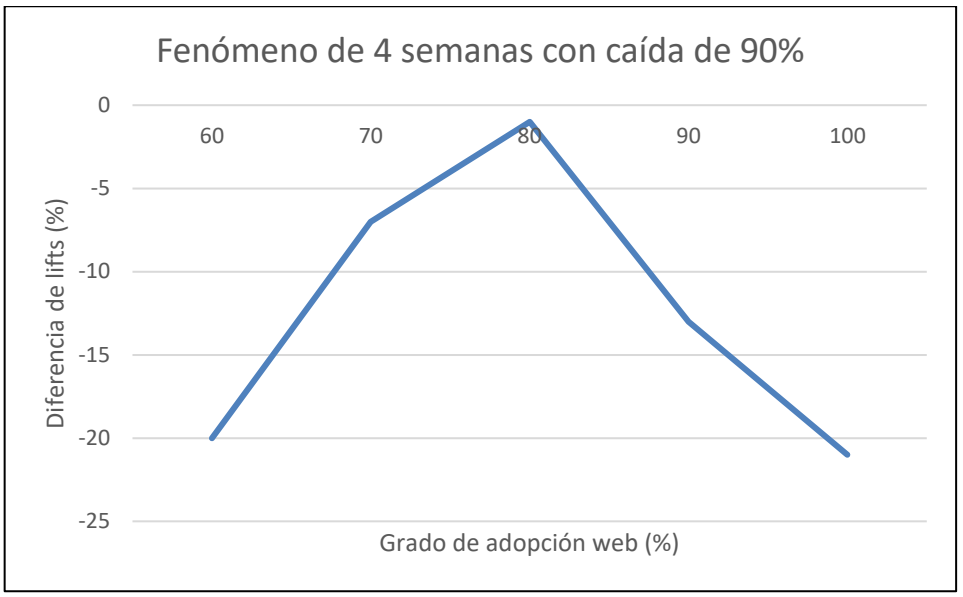

Ilustración 21. Definición del grado de adopción *web* para el fenómeno de 4 semanas y caída del 90% en la tasa.
Fuente: Elaboración propia.

De las figuras se observa que cuando se toman tres semanas, el grado de adopción *web* de un 60% presenta un *lift* positivo, lo que no ocurre cuando se consideran cuatro semanas. Esto significa que tres semanas no son un tiempo suficiente para personas con un menor grado de adopción *web*.

También se ve que a partir de una adopción del 80% existe, por un lado, un efecto en las transacciones por saturarse, dado que el *lift* es negativo, y además se tiene una relación entre el grado de adopción y el impacto en las transacciones.

Esta relación consiste en que a mayor grado de adopción *web*, mayor impacto transaccional. Por esta razón es que se escoge un 80% de adopción y no un 70%. No se incluye en los gráficos grados de adopción menores ya que sus *lift* de transacciones son cercanos o mayores a 0.

En otras palabras, mientras más *web* es el cliente, mayor es el impacto transaccional cuando se satura. El porcentaje de clientes que son al menos 80% *web* en el segmento alto es de un 6,64%.

ii. Definición del criterio de saturación

Los escenarios se generan a partir de una combinación de la caída en la tasa y la duración. A continuación, se muestran los escenarios candidatos a ser el fenómeno en el mes de marzo de 2018, escogido usando el indicador de impacto mencionado

en la metodología:

Duración	Caída tasa	Indicador marzo	Indicador mayo	Promedio indicadores
1	60%	-30,66	4,27	-13,19
3	60%	-8,87	-4,08	-6,47
1	50%	-9,35	3,98	-2,68
4	40%	-0,19	-4,79	-2,49
4	60%	-2,82	-1,92	-2,37
3	40%	-0,17	-4,26	-2,22

Tabla 8. Escenarios candidatos a ser el fenómeno escogido.
Fuente: Elaboración propia.

Se destaca de la tabla anterior que el indicador de mayo es positivo para los fenómenos que duran una semana. Esto muestra que esa cantidad de tiempo es insuficiente para establecer un impacto transaccional. La razón por la cual aparecen como candidatos es que cuentan con más personas que en los otros escenarios, que es uno de los factores del indicador.

Con esto, resta saber si estos impactos son estadísticamente significativos. Para eso se usa la regresión descrita en la metodología y estos son los resultados:

Fenómeno	p-valor Marzo	p-valor Mayo
1 semana con caída de 40%	0,4%	92,3%
3 semanas con caída de 40%	14%	14,1%
1 semana con caída de 50%	0,7%	36,7%
4 semanas con caída de 60%	6,4%	6,7%
4 semanas con caída de 40%	6,4%	98,5%
3 semanas con caída de 60%	3,9%	62,5%

Tabla 9. P-valores asociados a los escenarios candidatos.
Fuente: Elaboración propia.

A la vista de los resultados, se tiene que el fenómeno escogido es el de 4 semanas con una caída de un 60%. El coeficiente asociado a este fenómeno en marzo es de -0,14 y de -0,12 en mayo.

6.3.2 Segmento medio y bajo

Para estos segmentos, se observa que no existe un impacto transaccional producto de la saturación, ya que las diferencias de *lifts* son no negativas. Además, tampoco se observa siquiera que mientras más tiempo se mantiene una caída, menor diferencia en los *lifts* de transacciones, como sí ocurre en el segmento alto. A continuación, se muestran las diferencias en los *lifts* para los clientes con adopción *web* de un 80% para distintos escenarios y segmentos:

Segmento medio-marzo				
Caída\Semanas	1	2	3	4
70	2%	-1%	3%	5%
80	1%	2%	4%	6%
90	1%	4%	6%	8%

Tabla 10. Impacto transaccional en el mes de marzo para el segmento medio con adopción *web* de un 80%.
Fuente: Elaboración propia.

Segmento medio-mayo				
Caída\Semanas	1	2	3	4
70	3%	4%	-2%	5%
80	3%	2%	2%	6%
90	3%	3%	2%	7%

Tabla 11. Impacto transaccional en el mes de mayo para el segmento medio con adopción *web* de un 80%.
Fuente: Elaboración propia.

Segmento bajo-marzo y mayo				
Mes\Semanas	1	2	3	4
Marzo	-5%	7%	7%	9%
Mayo	3%	6%	7%	5%

Tabla 12. Impacto transaccional para los meses de marzo y mayo para el segmento bajo con adopción *web* de un 80%.
Fuente: Elaboración propia.

Por lo tanto, se descarta modelar la probabilidad de saturación para estos segmentos ya que se concluye que no hay evidencia de saturación.

6.4 Modelo predictivo de saturación

6.4.1 Generación de variable dependiente

Dado que se encuentra definido el fenómeno de saturación para el segmento alto, se debe generar esta variable. Ésta corresponde a una *dummy* que toma el valor 1 si es que la tasa de apertura de cuatro semanas posteriores al día en cuestión es menor o igual a un 40% de la tasa de apertura del mes previo, incluyendo el día evaluado.

Se utiliza información desde octubre de 2017 a julio de 2018. La primera fecha se debe a que la información usada para clasificar a los clientes según su adopción *web* se considera hasta septiembre de 2017. La segunda se debe a que se cuenta con información de *email marketing* hasta agosto de 2018, por tanto, se puede considerar hasta julio dado que se necesita un mes posterior para establecer si existe saturación.

Dado que existen datos previos al período de octubre de 2017 a julio de 2018, existe la posibilidad de que el primer registro de un cliente indique que éste se encuentra saturado. Estos casos se excluyen ya que se busca entender qué lleva un cliente que no está saturado a estarlo. También se excluyen los registros posteriores a la primera saturación del período estudiado, dado que considerar una saturación posterior a la primera para un cliente dado puede alterar las conclusiones del estudio.

6.4.2 Selección de variables con SCAD

El método de selección de variables, llamado SCAD, indica qué variable se puede descartar del modelo al hacer una estimación sobre el total de variables y asignando un valor no positivo al coeficiente asociado a la variable que se deba descartar. Los resultados detallados de la aplicación de este método a las variables utilizadas se encuentran en anexos (se encuentran los nombres originales de las variables, cuyo entendimiento es intuitivo). A continuación, se muestran las variables que este método considera:

Variables método SCAD	
Vistas a productos en *web* de negocio *retail* 12 meses	*Emails* abiertos últimos 2 días negocio *retail*
Clicks mismo día negocio *retail*	*Emails* abiertos últimos 3 días negocio *retail*
Emails enviados entre 19 a 9 hrs. mismo día neg. *Retail*	*Emails* abiertos últimos 3 días tarj. crédito

Gasto negocio tarj. Crédito 1 mes	*Emails* abiertos últimos 3 días mej. Hogar
Gasto *web* negocio tarj. crédito 3 meses	N° de hijos entre 3 a 8 años mujer
Gasto negocio tarj. crédito 6 meses	N° de hijos mayores de 18 años hombre
Gasto negocio *retail* 6 meses	Soltero con hijo más pequeño < 6 años
Largo del asunto promedio del día	Dispositivo que recibe *email* es un computador
Clicks últimos 15 días mej. hogar	Semana
Clicks últimos 2 días tarj. crédito	Soltero sin hijos
Clicks últimos 5 días negocio *retail*	Trx *web* tarjeta crédito 6 meses
Emails enviados últimos 2 días tarj. crédito	Trx *web* negocio *retail* 12 meses
Emails enviados entre 19 y 9 hrs. últ. 3 días negocio *retail*	Desuscripción
Emails enviados último domingo negocio *retail*	Visitas *web* negocio *retail* 6 meses
Emails enviados último domingo tarj. crédito	

Tabla 13. Variables consideradas por el método de selección de variables SCAD.
Fuente: Elaboración propia.

6.4.3 Selección del modelo

Se busca el modelo con mejor ajuste usando el método *backward*. Se cuenta con el modelo inicial especificado en la sección de selección de variables con SCAD. Los resultados de los distintos modelos probados se especifican con detalle en anexos E. A continuación, se muestra una tabla resumen de los modelos probados con el valor del AIC correspondiente:

Paso	Variables que abandonan	Criterio de salida	AIC
0	-	-	1834,6
1	Gasto tarj. Créd. 1 mes Desuscripción Trx *web* tarj. Créd. 6 meses Clicks últ. 5 días negocio *retail*	p-valor>90%	1832,1
2	*Emails* enviados últ. Domingo tarj. Créd. Soltero con hijo más pequeño < 6 años Vistas a productos en *web* de neg. *retail* 12 meses	p-valor>80%	1826,2
3	Clicks últimos 2 días tarj. Crédito N° de hijos mayores de 18 años hombre N° de hijos entre 3 a 8 años mujer	p-valor>50%	1821,1
4	*Emails* abiertos últimos 3 días tarj. Créd. Clicks mismo día negocio *retail* Largo del asunto promedio del día Gasto negocio tarj. crédito 6 meses *Emails* enviados últimos 2 días tarj. Crédito	p-valor>40%	1817,9
5	Clicks últimos 15 días mej. Hogar Gasto *web* negocio tarj. crédito 3 meses	p-valor>35%	1815,4
6	*Emails* enviados de 19 y 9 hrs. últ. 3 días neg.*retail* Soltero sin hijos	p-valor>15%	1815
7	*Emails* abiertos últimos 2 días negocio *retail* *Emails* abiertos últimos 3 días negocio *retail*	p-valor>5%	1813,8
8	Semana	p-valor>=4%	1816,2

Tabla 14. Iteraciones en la selección del modelo usando como indicador el AIC.
Fuente: Elaboración propia.

En cada iteración, usando el criterio de significancia, se excluyen las variables que cumplan con el criterio de salida especificado en la anterior. Se observa que en cada paso o iteración el valor del AIC va disminuyendo hasta llegar a la iteración 8 donde aumenta. Es justamente en esa iteración donde la variable excluida es significativa estadísticamente al 5%, valor definido en el marco teórico como referencia. Por aquello, no se continúa con las iteraciones.

Las variables consideradas luego del proceso anterior son las siguientes, donde se precisa respecto a la variable "*emails* enviados entre 19 a 9 hrs. mismo día neg. *retail*" que corresponde a recibir un *email* entre las 0 hrs. y 9 hrs. del mismo día donde se estudia la saturación o entre las 19 hrs. del mismo día donde se estudia la saturación y 0 hrs. del día siguiente, para así evitar confusiones. También se presenta el p-valor asociado:

Variable	p-valor
Semana	0,037
Visitas *web* negocio *retail* 6 meses	0,028
Emails abiertos últimos 3 días mej. hogar	0,025
Trx *web* negocio *retail* 12 meses	0,020
Gasto negocio *retail* 6 meses	0,012
*Email*s enviados último dgo. negocio *retail*	0,005
*Email*s enviados entre 19 a 9 hrs. mismo día neg. *retail*	0,003
Disp. que recibe *email* es un computador	0,000

Tabla 15. Variables consideradas por el método de *backward* selection.
Fuente: Elaboración propia.

6.4.4 Resolución de potenciales problemas de multicolinealidad

Se detectan potenciales problemas de multicolinealidad en el modelo a través del cálculo del factor de inflación de la varianza. La aplicación de este factor al modelo determinado anteriormente se resume en la siguiente tabla:

Variable	VIF
Disp. que usa es un pc	1,00
Emails enviados entre 19 a 9 hrs. mismo día neg. *retail*	1,03
Emails enviados último dgo. negocio *retail*	1,18

Gasto negocio *retail* 6 meses	1,02
Trx *web* negocio *retail* 12 meses	5,58
Emails abiertos últimos 3 días mej. hogar	1,03
Visitas *web* negocio *retail* 6 meses	5,56
Semana	1,21

Tabla 16. Detección de multicolinealidad a través del VIF aplicado al modelo original.
Fuente: Elaboración propia.

De la tabla, se tiene que la variable de transacciones *web* en el negocio de *retail* durante los últimos 12 meses presenta el VIF más alto y es mayor a 5, que fue definido previamente como valor crítico para detectar multicolinealidad. Por lo tanto, se elimina esta variable y se repite el cálculo del VIF.

Variable	VIF
Disp. que usa es un pc	1,00
Emails enviados entre 19 a 9 hrs. mismo día neg. *retail*	1,03
Emails enviados último dgo. negocio *retail*	1,18
Gasto negocio *retail* 6 meses	1,01
Emails abiertos últimos 3 días mej. hogar	1,03
Visitas *web* negocio *retail* 6 meses	1,02
Semana	1,18

Tabla 17. Detección de multicolinealidad a través del VIF aplicado al modelo modificado.
Fuente: Elaboración propia.

De la tabla se observa que todos los valores del VIF son cercanos a 1, por lo que se ha mitigado el problema de multicolinealidad. Además, se constata que el valor del AIC del modelo modificado es de 1810,2. Este valor es menor al del modelo original, que es de 1813,8. Al observar el p-valor de las variables en este nuevo modelo se observa lo siguiente:

Variable	p-valor
Semana	0,011
Visitas *web* negocio *retail* 6 meses	0,650
Emails abiertos últimos 3 días mej. hogar	0,014
Gasto negocio *retail* 6 meses	0,005
Emails enviados último dgo. negocio *retail*	0,053

*Email*s enviados entre 19 a 9 hrs. mismo día neg. *retail*	0,003
Disp. que recibe *email* es un computador	0,000

Tabla 18. P-valor de las variables posterior a tratamiento de multicolinealidad.
Fuente: Elaboración propia.

Se tiene que la variable de visitas *web* en el negocio de *retail* por los últimos 6 meses se aleja mucho de la significancia deseada, ya que probablemente al estar correlacionada con las transacciones *web*, se observa que conjuntamente son significativas, pero al excluir una de ellas la otra deja de ser significativa. Por lo tanto, se excluye del modelo.

6.4.5 Desempeño del modelo predictivo

Se *test*ea el desempeño del modelo predictivo especificado en la tabla anterior. Para esto, se calibra el modelo con distintas proporciones entre clases (saturado y no saturado) para conocer cuál es la configuración que aporta el mejor AUC sobre una muestra escogida al azar sin balancear. La tabla I muestra los valores del AUC sobre diferentes balanceos:

Balanceo (razón no saturados/saturados)	AUC
90/10	68,19%
80/20	67,07%
70/30	70,14%
60/40	71,64%
50/50	70,31%
40/60	72,48%
30/70	68,30%
20/80	80,30%
10/90	65,48%

Tabla 19. Valores del AUC de los distintos balanceos de la base de entrenamiento.
Fuente: Elaboración propia.

Se observa que la configuración que presenta mayor AUC y, por tanto, capacidad predictiva, es aquella calibrada con una base de entrenamiento compuesta por un 20% del caso mayoritario y un 80% del caso minoritario. El valor del AUC está por sobre un 80%, lo que se considera un fuerte efecto en términos predictivos[18].

[18] RICE, M. y HARRIS, G. 2005. Comparing effect sizes in follow-up studies: ROC Area,Cohen's d, and r. Law Hum Behav.29:615.

A continuación, se muestra la curva ROC resultante de la configuración 20/80:

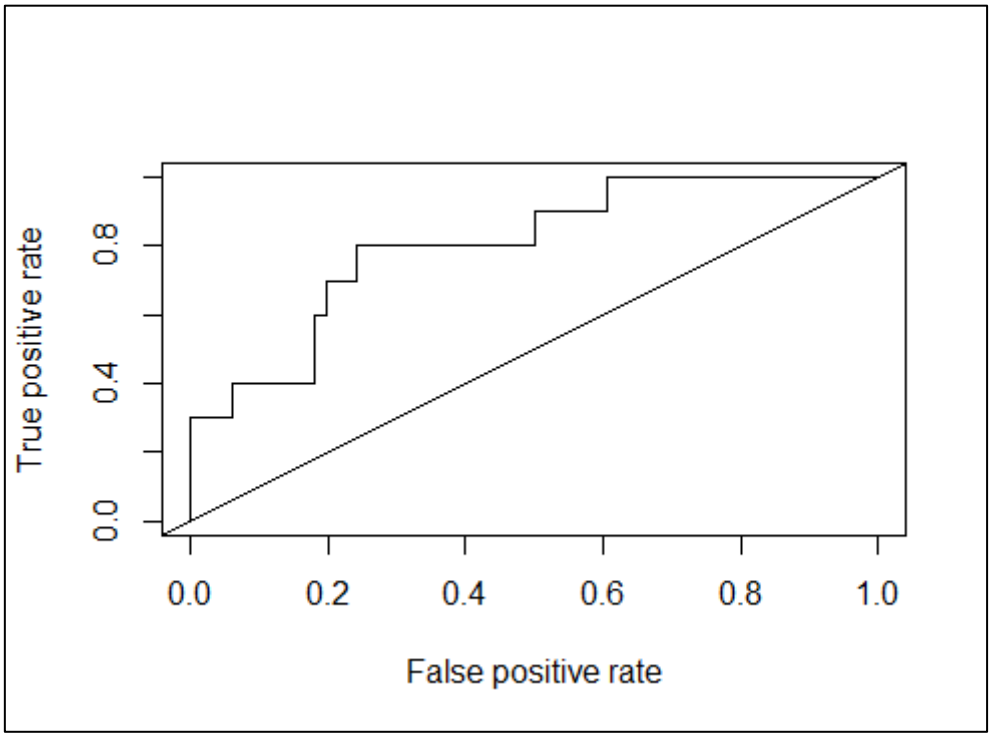

Ilustración 22. Curva ROC de la configuración de 80% casos saturados.
Fuente: RStudio.

6.4.6 Calibración del modelo predictivo

La calibración del modelo queda entonces como sigue, donde se debe recordar que la variable dependiente corresponde a una binaria que vale 1 si hay saturación y 0 si no:

Variable	Coeficiente estimado	Exp(coef. estimado)
Disp. que usa es un pc	0,35**	1,42
Emails enviados entre 19 a 9 hrs. mismo día neg. *retail*	-0,12**	0,88
Emails enviados último dgo. negocio *retail*	0,59*	1,81
Gasto negocio *retail* 6 meses	-0,00**	0,99
Emails abiertos últimos 3 días mej. hogar	-0,16*	0,85
Semana	-0,01*	0,99
Intercepto	1,65**	5,21

Tabla 20. Modelo predictivo calibrado.
(**): significativo al 1%; (*): significativo al 5%.
Fuente: Elaboración propia.

Dado lo explicado en la metodología respecto a la interpretación de los coeficientes de una regresión logística en términos de los *odds*, se tiene lo siguiente:

- Las variables de gasto en el negocio de *retail* durante los últimos 6 meses y semana poseen un valor de la función exponente sobre el coeficiente estimado cercano a 1. Esto habla de que no hay una asociación entre las variables, ni positiva ni negativa.

- Las variables de *email*s enviados entre las 19 y 9 hrs. por parte del negocio de *retail* del mismo día estudiado y los *email*s abiertos durante los últimos 3 días por parte del negocio de mejoramiento del hogar muestran una asociación negativa con la variable dependiente, por lo tanto, sería menos probable una saturación.

- Las variables dispositivo que usa el cliente es un computador y *email*s enviados el último domingo por parte del negocio de *retail* presentan una asociación positiva con la variable dependiente, por lo tanto, sería más probable una saturación.

6.5 Propuesta costo de envío

En la metodología se especifica los factores que influyen en el costo de envío. No es posible especificar con mayor detalle el costo dado que se desconoce lo que ocurre con el resto de los negocios, recordando que la idea de esto es generar incentivos a nivel de *holding* respecto a evitar la saturación. Dicho esto, se propone la siguiente formulación para el costo de envío:

$$Costo\ de\ envío\ = \text{Prob. Saturación}\ * \text{Coeficiente de adopción}\ web \tag{15}$$

Donde la probabilidad de saturación se calcula usando el modelo calibrado definido anteriormente. Por su parte, el coeficiente de adopción *web* se obtiene como sigue:

$$Coeficiente\ \text{de adopción}\ web = f(\gamma)$$
$$\gamma \in [0{,}8; 1] \tag{16}$$

Donde γ corresponde al grado de adopción de *web*, mientras que f corresponde a

una función creciente puesto que se observa que a mayor adopción *web* hay un mayor impacto transaccional.

VII. CONCLUSIONES

Existe bastante investigación acerca de los distintos factores que mejoran la tasa de apertura de los *emails*. La personalización, el diseño del *email*, el impacto del *trigger*, los tiempos de envío entre *email*, entre otros, han sido temas muy estudiados en memorias anteriores.

Sin embargo, no se ha estudiado en profundidad qué ocurre cuando existe un abuso de la herramienta que implique un deterioro en la relación con el cliente. Esta memoria busca abordar este tema, formalizando una definición de saturación y estudiando en qué tipo de clientes se puede establecer una asociación entre la saturación y una caída en términos transaccionales, dado que es algo que se puede cuantificar.

Para lograr esto, se desarrollan dos temas claves. El primero es la definición de la saturación, que es un concepto innovador dentro del estudio del *email marketing*. Este tema sirve de base como estudio para el negocio y además como insumo para la probabilidad de saturación. El segundo es el desarrollo de un modelo predictivo que permite determinar la probabilidad de saturación de un cliente un día dado a través del uso de métodos de selección de variables, métodos de balanceo y modelos probabilísticos.

7.1 Sobre el criterio de saturación

- Se detecta que para ningún segmento de clientes existe una relación general entre la saturación por *email marketing* y un cambio negativo en las transacciones. En otras palabras, se observa que en general las personas no dejan de comprar porque caiga su tasa de apertura de *emails*.

- Sin embargo, para una parte del segmento alto de clientes se encuentra que sí existe una relación entre saturarse y una caída en las transacciones. Estos clientes son aquellos que compran en el negocio principalmente usando el canal *web*.

- El porcentaje de adopción *web* a partir del cual se observa una clara relación entre la saturación y una baja en las transacciones es de un 80%. Es decir, aquellos clientes cuyas transacciones entre octubre de 2016 a septiembre de 2017, ambos meses incluidos, fueron compuestas por al menos un 80% de transacciones *web*, muestran una baja en sus transacciones al momento de saturarse.

- La definición de saturación para los clientes descritos en el párrafo anterior corresponde a presentar una caída en la tasa de un 60% durante cuatro semanas respecto a la tasa de referencia.

- Se observa que no existe una relación entre la saturación y el nivel de adopción *web* para el segmento medio ni para el segmento bajo. Esto es esperable a partir de lo estudiado por Ruth Rettie y Lisa Chittenden en su trabajo llamado *An evaluation of e-mail marketing and factors affecting response*, donde indican que el factor más relevante asociado a las altas tasas de apertura es la compra *online*. Por tanto, en estos segmentos donde la tasa de apertura no supera el 50% se espera que no haya una relación entre la apertura de *emails* con la compra *online*.

7.2 Sobre el modelo predictivo

- Se observa que las variables de semana y el gasto en el negocio de *retail* durante los últimos 6 meses muestran no impactar la probabilidad de saturación. Estas variables se consideran dado que hay un alto grado de seguridad sobre sus efectos.

- A partir de lo anterior, se observa que la saturación en el canal de *email marketing* no se ve afectada por variables transaccionales. Se tiene entonces que las transacciones *web* pueden ser impactadas por bajas en las tasas de apertura de *emails* pero no ocurre lo mismo en el sentido inverso, lo cual tiene sentido, ya que el correo es la comunicación constante del negocio con el cliente y la compra es la concretización de esta relación.

- Se muestra que las variables de cantidad de *emails* abiertos del negocio de mejoramiento del hogar los últimos 3 días y la cantidad de *emails* enviados entre 19 a 9 hrs. por parte del negocio de *retail* el mismo día implican una menor probabilidad de saturación.

- En el caso de la variable asociada al mejoramiento del hogar, se interpreta como un indicador de que el cliente abre constantemente los correos y no tiene intenciones de dejar de hacerlo, sobre todo considerando que en términos de imagen no existe una asociación entre este negocio y el negocio de *retail* en estudio, según el área donde se desarrolla el estudio.

- Por su parte, el resultado de la otra variable muestra que es una buena práctica enviar correos entre 19 a 9 hrs. a este tipo de clientes, lo que se condice con un estudio de Experian, una reconocida empresa que ofrece servicios de información y *marketing*[19]. Este estudio muestra que los clientes tienen una mejor recepción de los correos en términos de apertura entre las 20 hrs. y las 4 hrs. Probablemente esto se relaciona con que en ese rango de tiempo las personas revisan su correo personal dado que terminan su jornada laboral. Se sugiere estudiar rangos horarios contenidos dentro de las 19 hrs. a las 9 hrs. para entender con mayor certeza lo que ocurre.

- Se constata que las variables de cantidad de *emails* enviados durante el último domingo por parte del negocio de *retail* y si el cliente usa como dispositivo para abrir correos un computador implican una mayor probabilidad de saturación.

[19] Quarterly *Email* Benchmark Study. 2012. Experian *Marketing* Services. [20-01-2019]

- Respecto a la variable de *emails* recibidos el último domingo por parte del negocio de *retail*, se tiene como interpretación que este hecho causa molestia a los clientes, dado que el domingo es culturalmente un día de descanso y desconexión.

- Por su lado, la variable del dispositivo que abre el *email* muestra que los clientes que usan el computador habitualmente son más propensos a saturarse. Desde otra perspectiva, tiene sentido pensar que aquellos que lo hacen desde su celular son más flexibles respecto a la recepción de correos, dado que este dispositivo se diferencia del computador en que es de más fácil acceso. Por lo tanto, no se produce el acumulamiento de *emails*, dado que pueden revisarse al momento de ser recibidos. Dado este hallazgo, se sugiere identificar con información histórica si los clientes usan un celular o un computador para abrir sus *emails*.

- A partir de las distintas configuraciones de balance de los datos de entrenamiento se obtiene que la razón que aporta una mayor capacidad predictiva está compuesta por un 20% del caso mayoritario y un 80% del caso minoritario. Esto se entiende porque el modelo se nutre de más información acerca del fenómeno de saturación y por tanto puede detectarlo de mejor manera.

- El valor del AUC se puede interpretar como el porcentaje de las veces en que al elegir al azar un caso positivo (saturarse) éste es clasificado con un valor de probabilidad mayor que un caso negativo. Esto ocurre cerca de un 80% de las veces, lo que habla de una buena capacidad predictiva del modelo.

- El costo de envío debe considerar la adopción *web* de los clientes, dado que aquellos con mayor adopción a ese canal se ven más impactados transaccionalmente cuando bajan abruptamente su tasa de apertura. El negocio debe velar por cuidar a sus clientes *web* cuidando la manera en que se comunica con ellos, dado que se observa que son muy sensibles al *email marketing*.

7.3 Limitaciones y trabajos futuros

Una limitación es que existan factores que pudieran alterar el entendimiento del comportamiento de las personas respecto a la apertura de *emails*, tales como problemas con su correo en un determinado momento o cualquier otro evento desafortunado que no se puede controlar.

Como trabajo futuro se propone controlar eventos especiales como el *cyber day*, dado que tiene un comportamiento de compra y de apertura de *emails* distinto al resto del año. También se propone profundizar con un estudio experimental donde se controla la cantidad de toques.

Respecto a la segmentación, se propone generar segmentos que incluyan la dimensión de las transacciones. Es decir, tener segmentos de clientes con altas

tasas de apertura y distintos niveles de transacciones, de manera de entender si se pueden obtener conclusiones a partir de ahí.

Siguiendo con el criterio de saturación, se puede estudiar el efecto de la saturación a largo plazo, es decir, a más de tres meses. Con esto, se tiene la opción de estudiar conjuntamente qué ocurre con las tasas de apertura y las transacciones *web*. ¿Se alcanzan los niveles iniciales de apertura de *email*s? ¿De ocurrir esto, aumentan la cantidad de transacciones *web*? Éstas son el tipo de preguntas que se pueden responder.

Para el modelo de predicción se puede probar interactuar las variables buscando algún efecto significativo. Por ejemplo, crear la interacción entre una variable de transacciones *web* con otra del tipo sociodemográfica que se intuya pueda aportar a construir un perfil de los clientes *web*.

Para el costo de envío existe un gran espacio de mejora, en tanto se incorpore a este análisis una definición de saturación para los clientes del resto de los negocios del *holding*. Con esta información existen alternativas tales como optimizar alguna métrica que permita asignar valores conocidos a los costos de envío.

VIII. GLOSARIO

A lo largo del informe se utilizan términos que no resultan ser intuitivos necesariamente. En este apartado se procede a detallar los más relevantes:

- Visitas: una visita corresponde a una combinación única de un día de compra con un local. Es decir, si una persona compra muchas veces en un mismo local un día dado, esto equivale a una visita.

- Transacción: una transacción corresponde a una combinación única de una boleta, un día de compra y un local. Es decir, si una persona compra muchos productos de una sola vez, esto equivale a una transacción. Si compra muchos productos en distintos momentos, se tiene entonces más de una boleta y por ende más de una transacción.

- Escenario: en este contexto, se refiere a las distintas combinaciones de caídas en la tasa con duración de esta caída.

- *Cyber*: consiste en un evento durante el cual distintas empresas ofrecen productos con descuentos a través del canal *online*.

- ROI: conceptualmente, corresponde a una medida del rendimiento de una empresa en materia financiera.

- *Holding*: es una sociedad que administra un conjunto de empresas.

- *Trigger*: en *email marketing*, consiste disparar mensajes o notificaciones en un momento determinado, como respuesta a un evento.

- *Dummy*: es una variable que toma el valor de 0 ó 1 para indicar la ausencia o presencia, respectivamente, de algún hecho.

IX. BIBLIOGRAFÍA

1.- Regresión logística [en línea] <https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/regression/idh_lreg.html> [consulta: 10 Noviembre 2019]

2.- PRIORE, T. 2000. The Fall, Rise of E-Mail Response Rates [En línea] Channel Marketing <https://www.dmnews.com/channelmarketing/news/13094364/the-fall-rise-of-email-response-rates> [Consulta: 12 Noviembre 2019]

3.- GAUZENTE, C., RANCHHOD, A. and GURAU, C. 2008. SMS-marketing: a study of consumer saturation using an extended TAM approach. International Journal of Electronic Business, 6 (3), 282. (doi:10.1504/IJEB.2008.019108)

4.- LESKOVEC, J., ADAMIC, L. A., and HUBERMAN, B. A. 2007. The dynamics of viral marketing. ACM Trans. Web, 1, 1, Article 5 (May 2007), 39 pages. DOI = 10.1145/1232722.1232727 http://doi.acm.org/ 10.1145/1232722.1232727

5.- NIZAMUDDIN, I. Social Media Advertising In Malaysia: The Power Of Viral Marketing. International Journal of Business and Management Invention (IJBMI), vol. 07, no. 05, 2018, pp. 74–78.

6.- MACQUEEN, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bSMSp/1200512992

7.- STANTON, W., ETZEL, M. y WALKER B. 2007. Fundamentos de marketing. 14a ed. México, D. F.: McGraw Hill. xxv, 741 p. ISBN 9789701062012.

8.- CAMPOS, A. 2016. Estudio de respuestas de clientes frente a envío de emails automatizados en una tienda de retail. Memoria para optar al título de Ingeniero Civil Industrial. Universidad de Chile.

9.- FREEDMAN, D. 2009. Statistical Models: Theory and Practice. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815867

10.- WASSERSTEIN, R. and LAZAR, N. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

11.- GONZÁLEZ, A.2015. Selección de variables. Una revisión de métodos existentes. Máster en técnicas estadísticas. Universidade da Coruña. Facultad de informática.

12.- AKAIKE, H. 1974. A new look at the statistical model identification, IEEE Transactions on Automatic Control.

13.- TRAIN, K. 2009. Discrete Choice Methods with Simulation. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511805271

14.- COX, D. 1958. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society Series B, 20, 215-242.

15.- CÁRDENAS, J. 2015. Odd ratio: qué es y cómo se interpreta [En línea] < http://networkianos.com/odd-ratio-que-es-como-se-interpreta/#toc-1> [Consulta: 10 Enero 2019]

16.- FAYYAD, U., PIATETSKY-SHAPIRO, G. y SMYTH, P. 1996.The KDD Process for Extracting Useful Knowledge from Volumes of Data" Communications of the ACM, 39(11), 27-34.

17.- VELÁSQUEZ, L. y HITPASS, B. 2014. El nivel de Actividad en el Proceso Educativo como Indicador de Riesgo de Deserción Estudiantil medido en tiempo real con apoyo de tecnología BAM. Conference: JCC2014, Workshop on BPM WBPM en Universidad Católica del Maule, Talca. (DOI: 10.13140/2.1.3217.8880).

18.- KETCHEN, D., SHOOK, C. 1996. The application of cluster analysis in Strategic Management Research: An analysis and critique. Strategic Management Journal. 17 (6): 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

19.- CHITTENDEN, L. y RETTIE, R. 2003. An evaluation of e-mail marketing and factors affecting response. Journal of Targeting, Measurement and Analysis for Marketing. 11. 10.1057/palgrave.jt.5740078.

20.- KARAGIANNOPOULOS, M., ANYFANTIS, D., KOTSIANTIS, S. y PINTELAS, P. 2007. A Wrapper for Reweighting Training Instances for Handling Imbalanced Data Sets. In IFIP International Federation for Information Processing,Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C, Pnevmatikakis, L., Polymenakos, L., (Boston: Springer), pp. 29-36.

21.- CAMPOS, A. 2016. Estudio de respuestas de clientes frente a envío de emails automatizados en una tienda de retail. Memoria para optar al título de Ingeniero Civil Industrial. Universidad de Chile.

22.- COVARRUBIAS, G. 2012.Construcción y validación de una metodología de seguimiento para modelos de regresión logística. Memoria para optar al título de Ingeniero Civil Industrial. Universidad de Chile.

23.- RICE, M. y HARRIS, G. 2005. Comparing effect sizes in follow-up studies: ROC Area,Cohen's d, and r. Law Hum Behav.29:615.

24.- JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. 2017. An Introduction to Statistical Learning (8th ed.). Springer Science+Business Media New York. pp. 101-102.

25.- SHEATHER, S. 2009. A modern approach to regression with R. New York, NY: Springer.

X. ANEXOS

10.1 Anexos A: Variables

Transaccionales	Relacionales
Gasto negocio *retail* 1 mes	Clicks mismo día negocio *retail*
Trx negocio *retail* 1 mes	Envíos mismo día negocio *retail*
N° meses de compra negocio *retail* 1 mes	Aperturas mismo día negocio *retail*
Visitas negocio *retail* 1 mes	*Emails* enviados entre 19 a 9 hrs. mismo día neg. *retail*
Gasto negocio *retail* 3 meses	*Emails* enviados últimos 2 días negocio *retail*
Trx negocio *retail* 3 meses	*Emails* enviados últimos 3 días negocio *retail*
N° meses de compra negocio *retail* 3 meses	*Emails* enviados últimos 5 días negocio *retail*
Visitas negocio *retail* 3 meses	*Emails* enviados últimos 7 días negocio *retail*
Gasto negocio *retail* 6 meses	*Emails* enviados últimos 15 días negocio *retail*
Trx negocio *retail* 6 meses	*Emails* enviados últimos 30 días negocio *retail*
N° meses de compra negocio *retail* 6 meses	*Emails* enviados últimos 60 días negocio *retail*
Visitas negocio *retail* 6 meses	*Emails* enviados último domingo negocio *retail*
Gasto negocio *retail* 12 meses	*Emails* enviados entre 19 a 9 hrs. últ. 3 días neg. *retail*
Trx negocio *retail* 12 meses	*Emails* enviados últimos 2 días tarj. créd.
N° meses de compra negocio *retail* 12 meses	*Emails* enviados últimos 3 días tarj. créd.
Visitas negocio *retail* 12 meses	*Emails* enviados últimos 5 días tarj. créd.
Recency negocio *retail*	*Emails* enviados últimos 7 días tarj. créd.
Gasto *web* en negocio *retail* 1 mes	*Emails* enviados últimos 15 días tarj. créd.
Trx *web* en negocio *retail* 1 mes	*Emails* enviados últimos 30 días tarj. créd.
N° meses de compra *web* negocio *retail* 1 mes	*Emails* enviados últimos 60 días tarj. créd.
Visitas *web* en negocio *retail* 1 mes	*Emails* enviados último domingo tarj. créd.
Gasto *web* en negocio *retail* 3 meses	*Emails* enviados entre 19 a 9 hrs. últ. 3 días tarj. créd.
Trx *web* en negocio *retail* 3 meses	*Emails* enviados últimos 2 días mej. hogar
N° meses de compra *web* negocio *retail* 3 meses	*Emails* enviados últimos 3 días mej. hogar
Visitas *web* negocio *retail* 3 meses	*Emails* enviados últimos 5 días mej. hogar
Gasto *web* negocio *retail* 6 meses	*Emails* enviados últimos 7 días mej. hogar
Trx *web* negocio *retail* 6 meses	*Emails* enviados últimos 15 días mej. hogar
N° meses de compra *web* negocio *retail* 6 meses	*Emails* enviados últimos 30 días mej. hogar
Visitas *web* negocio *retail* 6 meses	*Emails* enviados últimos 60 días mej. hogar
Gasto *web* negocio *retail* 12 meses	*Emails* enviados último domingo mej. hogar
Trx *web* negocio *retail* 12 meses	*Emails* enviados entre 19 a 9 hrs. últ. 3 días mej. hogar
N° meses de compra *web* negocio *retail* 12 meses	*Emails* enviados últimos 2 días supermercado
Visitas *web* negocio *retail* 12 meses	*Emails* enviados últimos 3 días supermercado
Recency *web* negocio *retail*	*Emails* enviados últimos 5 días supermercado
Gasto negocio tarjeta crédito 1 mes	*Emails* enviados últimos 7 días supermercado
Trx tarjeta crédito 1 mes	*Emails* enviados últimos 15 días supermercado
N° meses de compra tarjeta crédito 1 mes	*Emails* enviados últimos 30 días supermercado
Visitas tarjeta crédito 1 mes	*Emails* enviados últimos 60 días supermercado
Gasto tarjeta crédito 3 meses	*Emails* enviados último domingo supermercado
Trx tarjeta crédito 3 meses	*Emails* enviados entre 19 a 9 hrs. últ. 3 días super.
N° meses de compra tarjeta crédito 3 meses	Aperturas últimos 2 días negocio *retail*
Visitas tarjeta crédito 3 meses	Aperturas últimos 3 días negocio *retail*
Gasto tarjeta crédito 6 meses	Aperturas últimos 5 días negocio *retail*
Trx tarjeta crédito 6 meses	Aperturas últimos 7 días negocio *retail*
N° meses de compra tarjeta crédito 6 meses	Aperturas últimos 15 días negocio *retail*
Visitas tarjeta crédito 6 meses	Aperturas últimos 30 días negocio *retail*
Gasto tarjeta crédito 12 meses	Aperturas últimos 60 días negocio *retail*
Trx tarjeta crédito 12 meses	Aperturas últimos 2 días tarj. créd.
N° meses de compra tarjeta crédito 12 meses	Aperturas últimos 3 días tarj. créd.
Visitas tarjeta crédito 12 meses	Aperturas últimos 5 días tarj. créd.
Recency tarjeta crédito	Aperturas últimos 7 días tarj. créd.
Gasto *web* tarjeta crédito 1 mes	Aperturas últimos 15 días tarj. créd.
Trx *web* tarjeta crédito 1 mes	Aperturas últimos 30 días tarj. créd.

N° meses de compra *web* tarjeta crédito 1 mes
Visitas *web* en tarjeta crédito 1 mes
Gasto *web* en tarjeta crédito 3 meses
Trx *web* en tarjeta crédito 3 meses
N° meses de compra *web* tarjeta crédito 3 meses
Visitas *web* tarjeta crédito 3 meses
Gasto *web* tarjeta crédito 6 meses
Trx *web* tarjeta crédito 6 meses
N° meses de compra *web* tarjeta crédito 6 meses
Visitas *web* tarjeta crédito 6 meses
Gasto *web* tarjeta crédito 12 meses
Trx *web* tarjeta crédito 12 meses
N° meses de compra *web* tarjeta crédito 12 meses
Visitas *web* tarjeta crédito 12 meses
Recency *web* tarjeta crédito
Gasto mejoramiento hogar 1 mes
Trx mejoramiento hogar 1 mes
N° meses de compra mejoramiento hogar 1 mes
Visitas mejoramiento hogar 1 mes
Gasto mejoramiento hogar 3 meses
Trx mejoramiento hogar 3 meses
N° meses de compra mejoramiento hogar 3 meses
Visitas mejoramiento hogar 3 meses
Gasto mejoramiento hogar 6 meses
Trx mejoramiento hogar 6 meses
N° meses de compra mejoramiento hogar 6 meses
Visitas mejoramiento hogar 6 meses
Gasto mejoramiento hogar 12 meses
Trx mejoramiento hogar 12 meses
N° meses de compra mejoramiento hogar 12 meses
Visitas mejoramiento hogar 12 meses
Recency mejoramiento hogar
Gasto *web* en mejoramiento hogar 1 mes
Trx *web* en mejoramiento hogar 1 mes
N° meses de compra *web* mejoramiento hogar 1 mes
Visitas *web* en mejoramiento hogar 1 mes
Gasto *web* en mejoramiento hogar 3 meses
Trx *web* en mejoramiento hogar 3 meses
N° meses de compra *web* mej. hogar 3 meses
Visitas *web* mejoramiento hogar 3 meses
Gasto *web* mejoramiento hogar 6 meses
Trx *web* mejoramiento hogar 6 meses
N° meses de compra *web* mej. hogar 6 meses
Visitas *web* mejoramiento hogar 6 meses
Recency *web* mejoramiento hogar
Gasto supermercado 1 mes
Trx supermercado 1 mes
N° meses de compra supermercado 1 mes
Visitas supermercado 1 mes
Gasto supermercado 3 meses
Trx supermercado 3 meses
N° meses de compra supermercado 3 meses
Visitas supermercado 3 meses
Gasto supermercado 6 meses
Trx supermercado 6 meses
N° meses de compra supermercado 6 meses
Visitas supermercado 6 meses
Gasto supermercado 12 meses
Trx supermercado 12 meses
N° meses de compra supermercado 12 meses
Visitas supermercado 12 meses

Navegación *web*

Agrega a bolsa en *web* de negocio *retail* 1 mes
Vistas a productos en *web* de negocio *retail* 1 mes
Cantidad de órdenes en *web* de negocio *retail* 1 mes
Aperturas últimos 60 días tarj. créd.
Aperturas últimos 2 días mej. hogar
Aperturas últimos 3 días mej. hogar
Aperturas últimos 5 días mej. hogar
Aperturas últimos 7 días mej. hogar
Aperturas últimos 15 días mej. hogar
Aperturas últimos 30 días mej. hogar
Aperturas últimos 60 días mej. hogar
Aperturas últimos 2 días supermercado
Aperturas últimos 3 días supermercado
Aperturas últimos 5 días supermercado
Aperturas últimos 7 días supermercado
Aperturas últimos 15 días supermercado
Aperturas últimos 30 días supermercado
Aperturas últimos 60 días supermercado
Clicks últimos 2 días negocio *retail*
Clicks últimos 3 días negocio *retail*
Clicks últimos 5 días negocio *retail*
Clicks últimos 7 días negocio *retail*
Clicks últimos 15 días negocio *retail*
Clicks últimos 30 días negocio *retail*
Clicks últimos 60 días negocio *retail*
Clicks últimos 2 días tarj. créd.
Clicks últimos 3 días tarj. créd.
Clicks últimos 5 días tarj. créd.
Clicks últimos 7 días tarj. créd.
Clicks últimos 15 días tarj. créd.
Clicks últimos 30 días tarj. créd.
Clicks últimos 60 días tarj. créd.
Clicks últimos 2 días mej. hogar
Clicks últimos 3 días mej. hogar
Clicks últimos 5 días mej. hogar
Clicks últimos 7 días mej. hogar
Clicks últimos 15 días mej. hogar
Clicks últimos 30 días mej. hogar
Clicks últimos 60 días mej. hogar
Clicks últimos 2 días supermercado
Clicks últimos 3 días supermercado
Clicks últimos 5 días supermercado
Clicks últimos 7 días supermercado
Clicks últimos 15 días supermercado
Clicks últimos 30 días supermercado
Clicks últimos 60 días supermercado

Sociodemográficas

Edad
N° de hijos
N° de hijos bebés hombre
N° de hijos bebés mujer
N° de hijos entre 3 a 8 años hombre
N° de hijos entre 3 a 8 años mujer
N° de hijos entre 9 a 17 años hombre
N° de hijos entre 9 a 17 años mujer
N° de hijos mayores de 18 años hombre
N° de hijos mayores de 18 años mujer
N° de vehículos
Soltero sin hijos
Recién casado sin hijos
Casado sin hijos
Casado con hijo más pequeño < 6 años
Casado con hijo más pequeño de 6 y 17 años
Casado con hijo más pequeño de 17 y 30 años
Casado con hijo más pequeño > 30 años
Soltero con hijo más pequeño < 6 años
Soltero con hijo más pequeño de 6 y 17 años
Soltero con hijo más pequeño de 17 y 30 años

Agrega a bolsa en *web* de negocio *retail* 3 meses	Cupo tarjeta crédito negocio
Vistas a productos en *web* de negocio *retail* 3 meses	Género masculino
Cant. de órdenes en *web* de negocio *retail* 3 meses	Pertenece a la Región Metropolitana
Agrega a bolsa en *web* de negocio *retail* 12 meses	Casado
Vistas a productos en *web* de negocio *retail* 12 meses	Segmento de valor negocio *retail*
Cant. de órdenes en *web* de negocio *retail* 12 meses	Segmento de valor tarjeta de crédito
Mail	Temporal
Dispositivo que recibe *email* es un móvil	Semana
Dispositivo que recibe *email* es un computador	Día de la semana
Largo del asunto promedio del día	Día del año
Rebota	Mes
Desuscripción	

Anexo 1. Variables a disposición del estudio.

10.2 Anexos B: Escenarios sin filtrar transacciones *web*

Segmento alto-marzo				
Caída\Semanas	1	2	3	4
70	-2%	-6%	-4%	2%
80	-1%	-5%	-3%	2%
90	-1%	-5%	-3%	3%

Anexo 2. Resumen escenarios para segmento alto en marzo sin filtrar por tipo de transacción.

Segmento alto-abril				
Caída\Semanas	1	2	3	4
70	-9%	-13%	-14%	-12%
80	-8%	-11%	-13%	-12%
90	-8%	-10%	-13%	-12%

Anexo 3. Resumen escenarios para segmento alto en abril sin filtrar por tipo de transacción.

Segmento alto-mayo				
Caída\Semanas	1	2	3	4
70	1%	3%	3%	0%
80	2%	4%	4%	0%
90	2%	6%	5%	-1%

Anexo 4. Resumen escenarios para segmento alto en mayo sin filtrar por tipo de transacción.

Segmento medio-marzo				
Caída\Semanas	1	2	3	4
70	3%	4%	7%	2%
80	3%	4%	8%	3%
90	3%	4%	8%	3%

Anexo 5. Resumen escenarios para segmento medio en marzo sin filtrar por tipo de transacción.

Segmento medio-abril				
Caída\Semanas	1	2	3	4
70	-4%	-1%	0%	-5%
80	-3%	0%	1%	-5%
90	-3%	0%	1%	-4%

Anexo 6. Resumen escenarios para segmento medio en abril sin filtrar por tipo de transacción.

Segmento medio-mayo				
Caída\Semanas	1	2	3	4
70	3%	4%	6%	0%
80	4%	4%	6%	1%
90	4%	5%	6%	0%

Anexo 7. Resumen escenarios para segmento medio en mayo sin filtrar por tipo de transacción.

Segmento bajo-marzo, abril y mayo				
Mes\Semanas	1	2	3	4
Marzo	-10%	-7%	-8%	-2%
Abril	-12%	-12%	-8%	-5%
Mayo	-11%	-10%	-6%	-5%

Anexo 8. Resumen escenarios para segmento bajo sin filtrar por tipo de transacción.

10.3 Anexos C: Selección de variables

Variable	Coeficiente SCAD
semana	0,722
unsubs	0,426
n_*emails*_3_dias_19a9	0,164
n_*emails*_open_3_dias_tarjeta	0,148
largo_asunto_prom	0,127
pc	0,109
clicks	0,075
n_hijos_mas_18_hombre	0,071
gasto_*retail*_6m	0,070
padre_soltero_1	0,065
visita_*retail*_6m_*web*	0,064
n_*emails*_open_3_dias	0,063
enviados_19_9	0,054
n_*emails*_open_2_dias	0,048
n_*emails*_open_3_dias_hogar	0,048
n_clicks_15_dias_hogar	0,038
n_clicks_5_dias	0,036
gasto_tarjeta_3m_*web*	0,032
n_*emails*_2_dias_tarjeta	0,015
n_*emails*_dgo_tarjeta	0,014
n_*emails*_dgo	0,007
n_hijos_3_8_mujer	0,005
soltero_sin_hijos	0,003
gasto_tarjeta_6m	0,000
gasto_tarjeta_1m	0,000
n_clicks_2_dias_tarjeta	0,000
cant_vista_*retail*_12m	0,000
trx_*retail*_12m_*web*	0,000
trx_tarjeta_6m_*web*	0,000
mobile	0,000
bounces	0,000
envios	0,000
n_*emails*_2_dias	0,000
n_*emails*_3_dias	0,000
n_*emails*_7_dias	0,000
n_*emails*_15_dias	0,000
n_*emails*_30_dias	0,000

n_*emails*_60_dias	0,000
n_*emails*_5_dias_tarjeta	0,000
n_*emails*_7_dias_tarjeta	0,000
n_*emails*_15_dias_tarjeta	0,000
n_*emails*_30_dias_tarjeta	0,000
n_*emails*_60_dias_tarjeta	0,000
n_*emails*_3_dias_19a9_tarjeta	0,000
n_*emails*_3_dias_hogar	0,000
n_*emails*_5_dias_hogar	0,000
n_*emails*_7_dias_hogar	0,000
n_*emails*_15_dias_hogar	0,000
n_*emails*_30_dias_hogar	0,000
n_*emails*_60_dias_hogar	0,000
n_*emails*_dgo_hogar	0,000
n_*emails*_3_dias_19a9_hogar	0,000
n_*emails*_2_dias_super	0,000
n_*emails*_3_dias_super	0,000
n_*emails*_5_dias_super	0,000
n_*emails*_7_dias_super	0,000
n_*emails*_15_dias_super	0,000
n_*emails*_30_dias_super	0,000
n_*emails*_60_dias_super	0,000
n_*emails*_3_dias_19a9_super	0,000
n_*emails*_open_5_dias	0,000
n_*emails*_open_7_dias	0,000
n_*emails*_open_15_dias	0,000
n_*emails*_open_30_dias	0,000
n_*emails*_open_60_dias	0,000
n_*emails*_open_5_dias_tarjeta	0,000
n_*emails*_open_7_dias_tarjeta	0,000
n_*emails*_open_15_dias_tarjeta	0,000
n_*emails*_open_30_dias_tarjeta	0,000
n_*emails*_open_60_dias_tarjeta	0,000
n_*emails*_open_5_dias_hogar	0,000
n_*emails*_open_7_dias_hogar	0,000
n_*emails*_open_15_dias_hogar	0,000
n_*emails*_open_30_dias_hogar	0,000
n_*emails*_open_60_dias_hogar	0,000
n_*emails*_open_3_dias_super	0,000
n_*emails*_open_5_dias_super	0,000
n_*emails*_open_7_dias_super	0,000

n_*emails*_open_30_dias_super	0,000
n_*emails*_open_60_dias_super	0,000
n_clicks_2_dias	0,000
n_clicks_3_dias	0,000
n_clicks_7_dias	0,000
n_clicks_15_dias	0,000
n_clicks_30_dias	0,000
n_clicks_60_dias	0,000
n_clicks_3_dias_tarjeta	0,000
n_clicks_5_dias_tarjeta	0,000
n_clicks_7_dias_tarjeta	0,000
n_clicks_30_dias_tarjeta	0,000
n_clicks_60_dias_tarjeta	0,000
n_clicks_2_dias_hogar	0,000
n_clicks_5_dias_hogar	0,000
n_clicks_7_dias_hogar	0,000
n_clicks_30_dias_hogar	0,000
n_clicks_60_dias_hogar	0,000
n_clicks_2_dias_super	0,000
n_clicks_3_dias_super	0,000
n_clicks_7_dias_super	0,000
n_clicks_30_dias_super	0,000
n_clicks_60_dias_super	0,000
gasto_*retail*_1m	0,000
trx_*retail*_1m	0,000
visita_*retail*_1m	0,000
gasto_*retail*_3m	0,000
n_mes_compra_*retail*_3m	0,000
visita_*retail*_3m	0,000
trx_*retail*_6m	0,000
n_mes_compra_*retail*_6m	0,000
visita_*retail*_6m	0,000
n_mes_compra_*retail*_12m	0,000
visita_*retail*_12m	0,000
recency_*retail*	0,000
trx_*retail*_1m_*web*	0,000
n_mes_compra_*retail*_1m_*web*	0,000
visita_*retail*_1m_*web*	0,000
gasto_*retail*_3m_*web*	0,000
trx_*retail*_3m_*web*	0,000
n_mes_compra_*retail*_3m_*web*	0,000

visita_*retail*_3m_*web*	0,000
gasto_*retail*_6m_*web*	0,000
trx_*retail*_6m_*web*	0,000
n_mes_compra_*retail*_6m_*web*	0,000
gasto_*retail*_12m_*web*	0,000
n_mes_compra_*retail*_12m_*web*	0,000
recency_*retail*_*web*	0,000
trx_tarjeta_1m	0,000
n_mes_compra_tarjeta_1m	0,000
visita_tarjeta_1m	0,000
gasto_tarjeta_3m	0,000
trx_tarjeta_3m	0,000
n_mes_compra_tarjeta_3m	0,000
visita_tarjeta_3m	0,000
trx_tarjeta_6m	0,000
n_mes_compra_tarjeta_6m	0,000
visita_tarjeta_6m	0,000
gasto_tarjeta_12m	0,000
trx_tarjeta_12m	0,000
n_mes_compra_tarjeta_12m	0,000
visita_tarjeta_12m	0,000
recency_tarjeta	0,000
gasto_tarjeta_1m_*web*	0,000
trx_tarjeta_1m_*web*	0,000
n_mes_compra_tarjeta_1m_*web*	0,000
visita_tarjeta_1m_*web*	0,000
trx_tarjeta_3m_*web*	0,000
n_mes_compra_tarjeta_3m_*web*	0,000
visita_tarjeta_3m_*web*	0,000
gasto_tarjeta_6m_*web*	0,000
n_mes_compra_tarjeta_6m_*web*	0,000
visita_tarjeta_6m_*web*	0,000
gasto_tarjeta_12m_*web*	0,000
trx_tarjeta_12m_*web*	0,000
visita_tarjeta_12m_*web*	0,000
recency_tarjeta_*web*	0,000
cant_ordenes_*retail*_1m	0,000
cant_vista_*retail*_3m	0,000
cant_ordenes_*retail*_3m	0,000
ag_bolsa_*retail*_6m	0,000
cant_vista_*retail*_6m	0,000

cant_ordenes_*retail*_6m	0,000
ag_bolsa_*retail*_12m	0,000
cant_ordenes_*retail*_12m	0,000
edad	0,000
n_hijos	0,000
n_hijos_bebe_hombre	0,000
n_hijos_bebe_mujer	0,000
n_hijos_3_8_hombre	0,000
n_hijos_9_17_hombre	0,000
n_hijos_9_17_mujer	0,000
n_hijos_mas_18_mujer	0,000
n_vehiculos	0,000
casado_sin_hijos	0,000
nido_1	0,000
nido_2	0,000
nido_3	0,000
nido_vacio	0,000
padre_soltero_2	0,000
padre_soltero_3	0,000
cupo	0,000
rm_dum	0,000
seg_valor_tarjeta	0,000
seg_valor_fal	0,000
fen_filter	0,000
genero_dum	0,000
trx_*retail*_3m	0,000
trx_*retail*_12m	0,000
gasto_*retail*_1m_*web*	0,000
cant_vista_*retail*_1m	0,000
n_mes_compra_tarjeta_12m_*web*	-0,001
mes	-0,001
dia_anio	-0,001
n_mes_compra_*retail*_1m	-0,001
n_*emails*_open_15_dias_super	-0,002
n_*emails*_dgo_super	-0,003
dia	-0,003
n_*emails*_open_2_dias_super	-0,003
n_clicks_5_dias_super	-0,006
ag_bolsa_*retail*_3m	-0,011
n_*emails*_open_2_dias_tarjeta	-0,012
n_*emails*_open_2_dias_hogar	-0,018

recien_casado_sin_hijos	-0,022
n_clicks_15_dias_super	-0,025
gasto_*retail*_12m	-0,030
n_*emails*_2_dias_hogar	-0,034
visita_*retail*_12m_*web*	-0,044
n_*emails*_5_dias	-0,045
opens	-0,082
est_civil_dum	-0,096
n_clicks_15_dias_tarjeta	-0,106
n_*emails*_3_dias_tarjeta	-0,117
n_clicks_3_dias_hogar	-0,209

Anexo 9. Resultados del método SCAD sobre el total de las variables disponibles.

10.4 Anexos D: Selección del modelo

Variable	p-valor
gasto_tarjeta_1m	0,980
unsubs	0,962
trx_tarjeta_6m_*web*	0,934
n_clicks_5_dias	0,906
padre_soltero_1	0,872
cant_vista_*retail*_12m	0,813
n_*emails*_dgo_tarjeta	0,805
(Intercept)	0,752
n_hijos_3_8_mujer	0,627
n_hijos_mas_18_hombre	0,618
gasto_tarjeta_6m	0,593
clicks	0,543
n_*emails*_2_dias_tarjeta	0,522
n_*emails*_open_3_dias_tarjeta	0,485
largo_asunto_prom	0,484
n_clicks_2_dias_tarjeta	0,449
n_clicks_15_dias_hogar	0,359
gasto_tarjeta_3m_*web*	0,299
n_*emails*_3_dias_19a9	0,271
soltero_sin_hijos	0,208
n_*emails*_open_3_dias_hogar	0,090
n_*emails*_open_2_dias	0,069
n_*emails*_open_3_dias	0,058
trx_*retail*_12m_*web*	0,041
visita_*retail*_6m_*web*	0,036
semana	0,026
n_*emails*_dgo	0,019
gasto_*retail*_6m	0,007
enviados_19_9	0,004
pc	0,000

Anexo 10. Iteración 0 del método backward.

Variables	p-valor
n_*emails*_dgo_tarjeta	0,861
padre_soltero_1	0,855
cant_vista_*retail*_12m	0,823
(Intercept)	0,769
n_clicks_2_dias_tarjeta	0,744
n_hijos_mas_18_hombre	0,669
n_hijos_3_8_mujer	0,603
n_*emails*_open_3_dias_tarjeta	0,509
clicks	0,460
n_*emails*_2_dias_tarjeta	0,445
largo_asunto_prom	0,444
gasto_tarjeta_6m	0,425
gasto_tarjeta_3m_*web*	0,335
n_clicks_15_dias_hogar	0,315
n_*emails*_3_dias_19a9	0,295
soltero_sin_hijos	0,181
n_*emails*_open_3_dias_hogar	0,078
n_*emails*_open_2_dias	0,071
n_*emails*_open_3_dias	0,067
trx_*retail*_12m_*web*	0,036
visita_*retail*_6m_*web*	0,035
semana	0,028
n_*emails*_dgo	0,021
gasto_*retail*_6m	0,016
enviados_19_9	0,004
pc	0,000

Anexo 11. Iteración 1 del método backward.

Variable	p-valor
(Intercept)	0,776163
n_clicks_2_dias_tarjeta	0,747222
n_hijos_mas_18_hombre	0,646725
n_hijos_3_8_mujer	0,537106
n_*emails*_open_3_dias_tarjeta	0,495856
clicks	0,466846
largo_asunto_prom	0,453045
n_*emails*_2_dias_tarjeta	0,433229
gasto_tarjeta_6m	0,431356
gasto_tarjeta_3m_*web*	0,336227
n_clicks_15_dias_hogar	0,315645
n_*emails*_3_dias_19a9	0,300182
soltero_sin_hijos	0,181931
n_*emails*_open_3_dias_hogar	0,08032
n_*emails*_open_2_dias	0,069062
n_*emails*_open_3_dias	0,066233
visita_*retail*_6m_*web*	0,036192
trx_*retail*_12m_*web*	0,031702
semana	0,025414
n_*emails*_dgo	0,019722
gasto_*retail*_6m	0,016291
enviados_19_9	0,003888
pc	7,69E-05

Anexo 12. Iteración 2 del método backward.

Variable	p-valor
(Intercept)	0,730
n_*emails*_open_3_dias_tarjeta	0,515
clicks	0,512
largo_asunto_prom	0,436
gasto_tarjeta_6m	0,436
n_*emails*_2_dias_tarjeta	0,407
gasto_tarjeta_3m_*web*	0,331
n_clicks_15_dias_hogar	0,323
n_*emails*_3_dias_19a9	0,295
soltero_sin_hijos	0,127
n_*emails*_open_3_dias_hogar	0,083
n_*emails*_open_2_dias	0,063
n_*emails*_open_3_dias	0,060
visita_*retail*_6m_*web*	0,041
trx_*retail*_12m_*web*	0,034
semana	0,026
n_*emails*_dgo	0,019
gasto_*retail*_6m	0,017
enviados_19_9	0,004
pc	0,000

Anexo 13. Iteración 3 del método backward.

Variable	p-valor
(Intercept)	0,532
n_clicks_15_dias_hogar	0,406
gasto_tarjeta_3m_*web*	0,366
n_*emails*_3_dias_19a9	0,236
soltero_sin_hijos	0,150
n_*emails*_open_2_dias	0,070
n_*emails*_open_3_dias	0,063
n_*emails*_open_3_dias_hogar	0,059
visita_*retail*_6m_*web*	0,032
semana	0,024
trx_*retail*_12m_*web*	0,022
gasto_*retail*_6m	0,009
n_*emails*_dgo	0,009
enviados_19_9	0,001
pc	0,000

Anexo 14. Iteración 4 del método backward.

Variable	p-valor
(Intercept)	0,532
n_*emails*_3_dias_19a9	0,209
soltero_sin_hijos	0,151
n_*emails*_open_2_dias	0,083
n_*emails*_open_3_dias	0,073
n_*emails*_open_3_dias_hogar	0,048
semana	0,025
visita_*retail*_6m_*web*	0,025
trx_*retail*_12m_*web*	0,021
gasto_*retail*_6m	0,011
n_*emails*_dgo	0,008
enviados_19_9	0,001
pc	0,000

Anexo 15. Iteración 5 del método backward.

Variable	p-valor
(Intercept)	0,258
n_*emails*_open_3_dias	0,121
n_*emails*_open_2_dias	0,100
n_*emails*_open_3_dias_hogar	0,056
semana	0,038
visita_*retail*_6m_*web*	0,027
trx_*retail*_12m_*web*	0,020
gasto_*retail*_6m	0,013
n_*emails*_dgo	0,005
enviados_19_9	0,002
pc	0,000

Anexo 16. Iteración 6 del método backward.

Variable	p-valor
(Intercept)	0,224
semana	0,037
visita_*retail*_6m_*web*	0,028
n_*emails*_open_3_dias_hogar	0,025
trx_*retail*_12m_*web*	0,020
gasto_*retail*_6m	0,012
n_*emails*_dgo	0,005
enviados_19_9	0,003
pc	0,000

Anexo 17. Iteración 7 del método backward.

Variable	p-valor
(Intercept)	0,002
pc	0,000
gasto_*retail*_6m	0,020
visita_*retail*_6m_*web*	0,043
enviados_19_9	0,002
n_*emails*_open_3_dias_hogar	0,024
n_*emails*_dgo	0,035
trx_*retail*_12m_*web*	0,035

Anexo 18. Iteración 8 del método backward.