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

DISEÑO DE UNA POLÍTICA DE TOQUES EN EMAIL MARKETING DIFERENCIADA POR CLIENTES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JUAN JOSÉ ESCOBEDO OLAVARRÍA

PROFESOR GUÍA:
CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
JUAN PABLO ROMERO GODOY

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Juan José Escobedo Olavarría
FECHA: 2022
PROFESOR GUÍA: Carolina Segovia Riquelme

DISEÑO DE UNA POLÍTICA DE TOQUES EN EMAIL MARKETING DIFERENCIADA POR CLIENTES

El email marketing es una forma comunicacional que utilizan la mayoría de las empresas a lo largo del mundo para ofrecer productos o mantener el contacto con los clientes. Este mecanismo permite alcanzar grandes números de personas a través del envío masivo de mails. No obstante, esta técnica puede perjudicar el grado de recepción que tengan los individuos, quienes podrían querer recibir diferentes cantidades de emails o utilizar distintos horarios para revisar sus correos electrónicos. Ante esto, el proyecto de titulación esbozado en este informe busca resolver esta problemática, la cual afecta al área CRM Personas de un banco chileno.

El CRM Personas lleva a cabo un proceso de envío de correos que comienza con la selección y priorización del público a contactar, siguiendo con la preparación de los emails para finalmente realizar los envíos y obtener una respuesta. Si bien la empresa establece ciertas restricciones para el envío de mails, estas se caracterizan por ser las mismas para todos los clientes. Por esto, el objetivo de este trabajo es diseñar una política de envío de correos en que se personalice el número de mails a mandar y el horario de envío de estos, según el segmento al que pertenezca la persona contactada. De esta forma se busca optimizar las tasas de lectura (o apertura) y la desuscripción de las campañas de email marketing que desarrolla la organización.

Para cumplir este objetivo, se define un criterio de saturación en email que permite examinar el grado de rechazo que tiene este medio por parte de las personas. Luego, se determinan 5 segmentos de clientes: Baja lectura, Saturados que no leen, Alta lectura, Saturados que leen y Mediana lectura; los cuales promedian tasas de lecturas cercanas a 15%, 0%, 90%, 25% y 50% respectivamente. Además, se determina un límite de correos que los individuos pueden recibir en un periodo de tiempo: entre 1 o 22 mails por mes, según el segmento que integre la persona. Posteriormente se establecen los horarios en que es más conveniente enviar los correos, de acuerdo al contenido del mismo.

Adicionalmente, se detallan algunos experimentos que permitirían probar los resultados de este trabajo, para así confirmar si estos ayudan a mejorar la recepción que tienen las personas del email marketing del banco.

Finalmente, se concluyen aspectos que demuestran que la empresa debiese implementar políticas de toques diferenciadas: las personas son capaces de leer distintas cantidades de correos mensualmente, y los mails debiesen ser enviados en diferentes días y horas (incluso podrían utilizarse horarios que el banco suele restringir, pero que muestran mejores efectos que aquellos que se usan comúnmente).

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. Introducción

El correo electrónico ha sido uno de los canales de comunicación más usados por empresas de diferentes rubros al caracterizarse por ser asequible y rentable. En particular, la organización con el que se está trabajando ve al email como un medio principal para contactarse con las personas, y dar a conocer sus estrategias de marketing. Sin embargo, la forma en que se realiza este email marketing en la empresa suele ser generalizada, es decir, no se diferencia entre clientes.

Ante esto, el presente trabajo de título busca diseñar políticas de envío de correos para que la empresa pueda ajustar aspectos como la cantidad de emails que se le mandan a una persona, o el día y la hora de envío de los correos, con el fin de maximizar la recepción de las campañas de email marketing de la organización.

1.1 Empresa

La organización con que se realiza este proyecto de titulación corresponde a un banco fundado en 1937 en Chile, el cual hasta el año 2021, ha establecido filiales en Estados Unidos y oficinas en México, Perú, Colombia, Brasil y China.

La misión de la organización se enuncia como: "El banco se define como una corporación de soluciones financieras que participa en todos los negocios y operaciones financieras que la Ley General de Bancos le permite, ofreciendo a la comunidad productos y servicios con procesos de alta eficiencia operacional y excelencia en la calidad, con una permanente innovación tecnológica, prudentes políticas de administración de riesgos y exigentes estándares éticos, los que deben ser respetados por todas las personas que se desempeñan en sus empresas. En este marco, y con el propósito de cumplir sus objetivos y políticas, la corporación se compromete a cuidar que dichos logros se obtengan con especial énfasis en los que considera sus cuatro pilares fundamentales" [22].

Regulado por la Comisión para el Mercado Financiero (CMF), el banco en cuestión atiende principalmente a tres tipos de clientes: Personas, Empresarios y Empresas. Para cada tipo de cliente, la organización ofrece productos y/o servicios tales como créditos, seguros, cuentas, entre otros. No obstante, el enfoque de este trabajo estará sobre el primer grupo, para los cuales, los productos y servicios ofrecidos son: cuentas corrientes y primas, tarjetas de crédito, créditos hipotecarios y de consumo, seguros, inversiones, y beneficios (descuentos, promociones, programas y viajes).

El año 2020, la organización tuvo una utilidad consolidada de más de $317 mil millones de pesos[1], significando una caída de 21,2% respecto de las utilidades del año 2019 (efecto que se atribuye al contexto de pandemia y a un mayor gasto para anticipar riesgos futuros e incrementar el ratio de cobertura) [22]. A pesar de esta disminución en la utilidad, destaca un alza en la participación de mercado en utilidades, la cual creció en cerca de 10 puntos porcentuales (p.p.). Esto refleja el

[1] A lo largo de este trabajo se utiliza como moneda el peso chileno, a menos que se especifique algo distinto.

impacto negativo que tuvo la pandemia por coronavirus que caracterizó al año 2020, la que no tan solo disminuyó las utilidades de esta organización, sino que lo hizo en todo el rubro bancario. Lo anteriormente descrito se puede observar de manera gráfica en la siguiente figura:

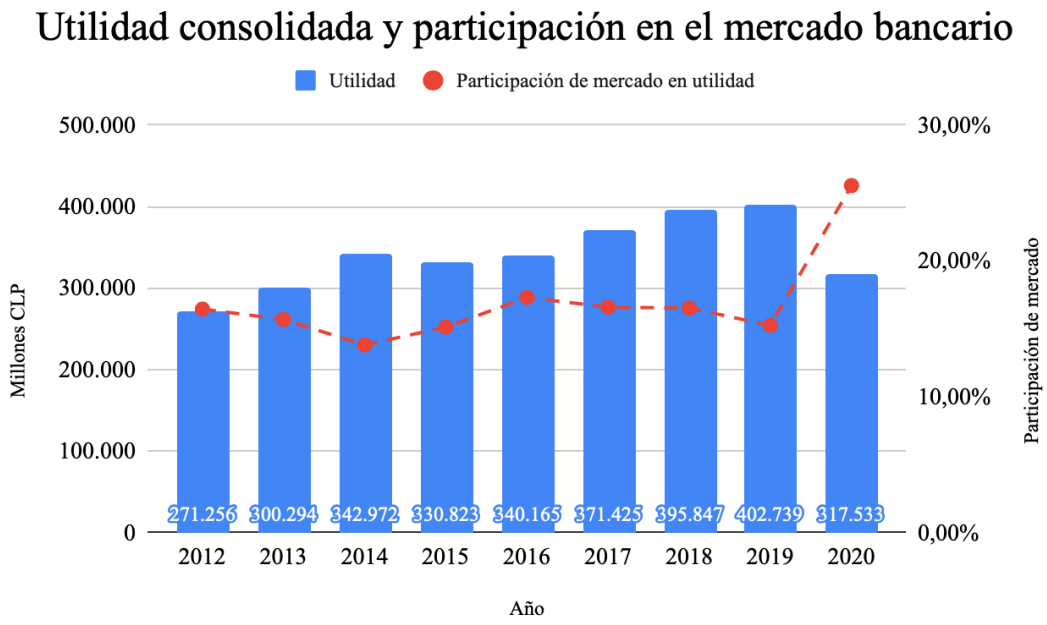

Figura 1: Utilidad consolidada y participación de mercado en utilidad de la empresa.
Fuente: Memoria Anual 2020 del banco.

Según una publicación en el sitio *Rankia*, donde se analizan cifras de la CMF, esta organización era la entidad bancaria con más activos totales administrados para el año 2021. Además, en febrero del mismo año, la revista *PaySpace Magazine* destacó a esta empresa entre los mejores bancos establecidos en Chile, mientras que la revista *Global Finance* posicionó a esta entidad dentro de los 10 bancos más seguros en América Latina el año 2020. De esta forma, es válido concluir que la organización se encuentra dentro de las entidades más importantes del mercado bancario chileno.

1.2 Información del área

El proyecto de titulación se realiza en la Gerencia de *Data y Analytics* (D&A) del banco, específicamente en el área de *Customer Relationship Management* de clientes Personas de la gerencia (CRM Personas); la cual se encarga de optimizar los resultados de venta, vinculación y atracción de clientes mediante la orquestación de *leads*[2] en los diferentes canales por los que el banco ofrece sus productos. En esa línea, esta área ofrece a los clientes tipo Personas los productos de la cartera respectiva, mediante las plataformas de distribución con las que cuentan, por ejemplo, *Salesforce*.

[2] *Lead* corresponde a cuando una empresa se contacta con un usuario que facilitó sus datos para poder ser comunicado.

El CRM Personas del banco está compuesto por un total de 11 trabajadores: un equipo de 10 *Business Analyst* más el gerente del área, donde este último es el solicitante de este trabajo de título.

El email marketing es uno de los métodos que utiliza el CRM Personas para comunicarse con los clientes de forma rápida y masiva. Según cifras extraídas desde las bases de datos del área, se llevan a cabo más de 200 campañas de email marketing por mes, lo que resulta en el envío de más de 10 millones de correos mensuales, demostrando la importancia que tiene este canal de comunicación para el área.

1.3 Email marketing

El email marketing (también conocido como *emailing*) es una herramienta de comunicación digital del que disponen las empresas para contactarse con sus clientes. Este mecanismo, define el envío de correos electrónicos a una base de contactos, entre los que se encuentran clientes vigentes y potenciales.

El *emailing* sirve, por ejemplo, para ofrecer productos a los clientes, informar de novedades o simplemente mantener el contacto, etc. Esta forma de comunicación ofrece a las empresas un retorno de la inversión (ROI) alto comparado con otros canales de distribución [11], considerando que el costo marginal por enviar un correo es despreciable y que los beneficios incluyen: conocer en profundidad al cliente, mejorar la calidad de atención o aumentar las ventas, entre otros. Según un estudio realizado por el *Content Marketing Institute* (CMI) en conjunto con *MarketingProfs*, el correo electrónico es de los canales preferidos por los expertos en marketing de contenidos B2C para distribuir contenido comercial. Más aún, 76% de los encuestados declaran usar este medio, posicionando así al email como el tercer canal de comunicación más usado para aquella distribución. El resultado de esta parte del estudio mencionado se puede ver en la figura a continuación:

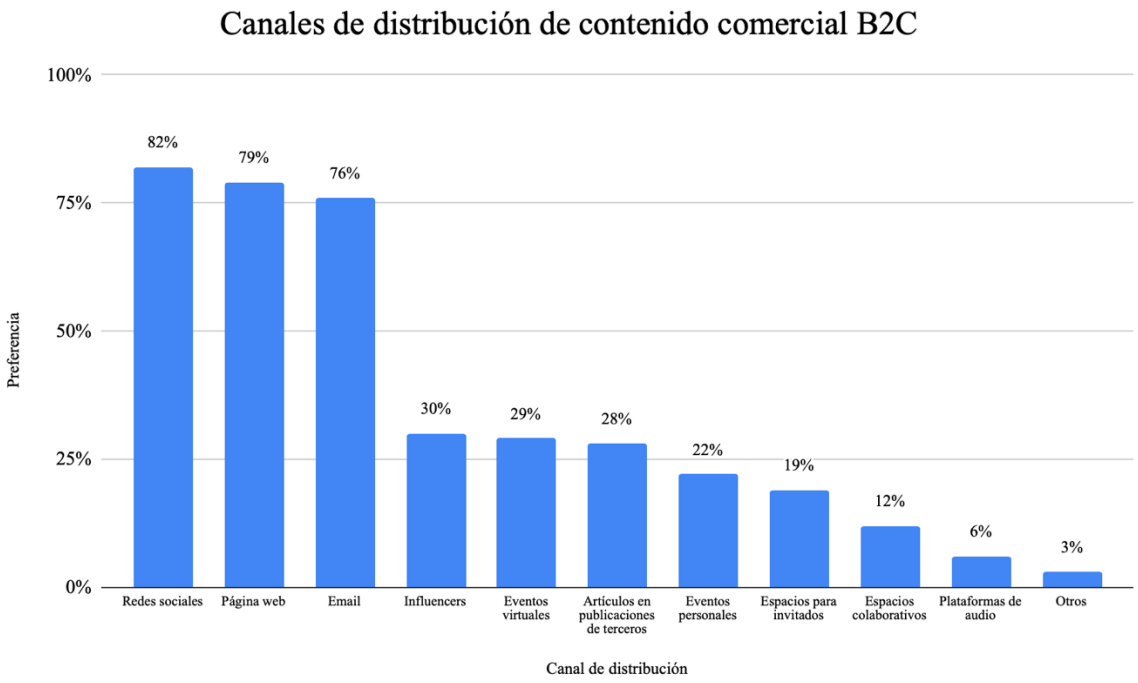

Figura 2: Canales de distribución de contenido comercial B2C utilizados el año 2020.
Fuente: "2021 *Content Marketing* B2C", CMI.

Cabe destacar que en el año 2020, más de 4 mil millones de personas usaron el correo electrónico [9], cifra que seguirá aumentando en el tiempo y que tiene sentido considerando el impacto que ha tenido en esta plataforma el creciente uso de *smartphones* y computadores, los cuales permiten acceder al email en cualquier lugar y momento. Esto ha provocado que las personas revisen su correo constantemente durante el día, significando incluso que esta acción sea un hábito en la rutina diaria de algunos individuos. Por ende, se incrementa el uso del email marketing por parte de las empresas para así alcanzar un mayor número de clientes a través de este medio.

2. Descripción del estudio

2.1 Problema

En general, el email marketing se distingue por utilizar diferentes mecanismos para lograr mayor atractivo hacia los clientes. Por ejemplo, existen casos en que se incorpora el nombre de una persona en el asunto del mail o se personaliza el cuerpo del correo. Si bien la organización ha incorporado aspectos como estos en sus emails, las campañas de *emailing* de la empresa también se caracterizan por generalizar variables como la cantidad de mails que se envían a las personas y los horarios de envío[3]. En otras palabras, no existen indicaciones que adapten el número de correos a enviar a un individuo en particular, o el día y la hora para realizar estos envíos.

[3] El horario de envío considera día y hora de envío.

En esa línea, al momento de configurar el email marketing, el banco no contempla la respuesta de los clientes frente a los correos que reciben. Esta respuesta se refleja en el *open rate*[4] de las campañas de *emailing*, el cual varía mensualmente entre 30% y 35% en promedio, según muestran los registros de los correos enviados el primer semestre del año 2021. En resumen, la política de envío[5] de correos que se utiliza con una persona que se caracteriza por leer muchos mails, es la misma que se ocupa con otra que abre pocos correos.

Finalmente, el gerente del CRM Personas señala que existe una correlación positiva entre la lectura de mails y la adquisición de productos por parte de las personas. Por esto, la organización busca establecer una política de toques[6] diferenciada por segmentos de clientes, la cual indique un número de contactos a establecer mediante email y un horario de envío de los correos, con tal de que mejoren las tasas de apertura y disminuya la desuscripción[7] del email marketing de la empresa.

2.2 Justificación del problema

El problema planteado anteriormente tiene dos aristas de justificación que se explicarán en esta parte: que el email marketing del banco evidencia una oferta no óptima, y la desuscripción de clientes.

2.2.1 Oferta no óptima

El no considerar, en las políticas de envío de correos, la respuesta que tienen los clientes a las campañas de email marketing, provoca que se envíen cantidades no óptimas de mails a las personas. En esa línea, se puede observar en las siguientes dos figuras la evolución en la tasa de apertura de un par de clientes, frente a la cantidad de correos que se les enviaron mensualmente:

[4] También tasa de apertura o tasa de lectura en español. Corresponde a un ratio que indica la fracción de correos electrónicos que son abiertos o leídos por los clientes contactados en un intervalo de tiempo, respecto del total de emails que se envían en el mismo intervalo. Se asume que las personas abren correos para leerlos, por lo que apertura y lectura se consideran sinónimos en este trabajo. Así, por ejemplo, si se mandan 100 correos en total, y de estos 50 son leídos, el *open rate* es 50%.
[5] Incluye cantidad de correos a enviar, días y horas de envío.
[6] Un toque corresponde al envío de un correo a un cliente.
[7] Corresponde a la acción con que una persona se auto elimina de las bases de contactos del banco, por ende, no se le debería contactar por el canal de comunicación que esa persona especifique.

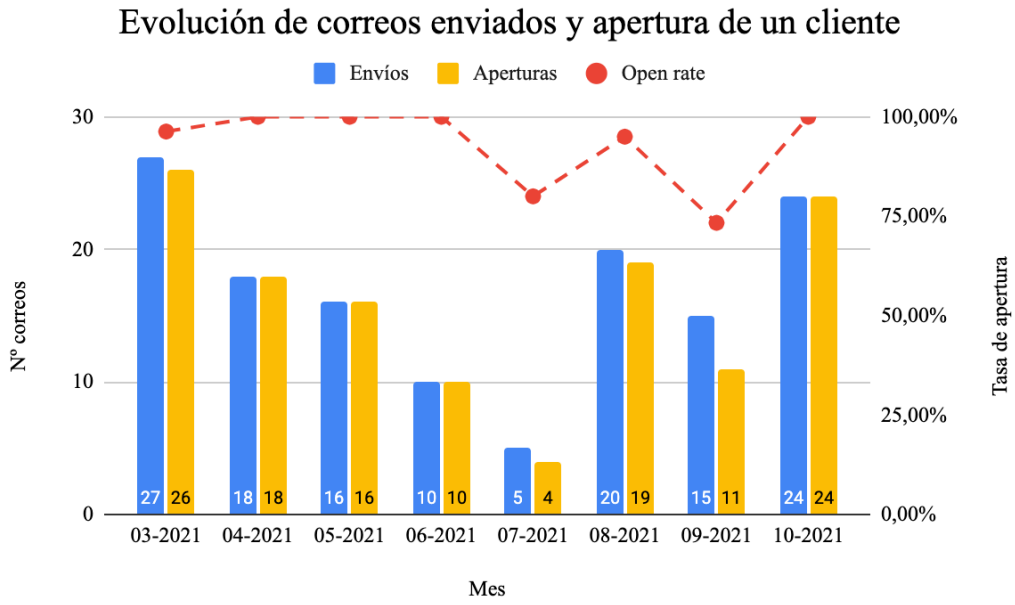

Figura 3: Cantidad de correos enviados y leídos por un cliente con alta tasa de apertura.
Fuente: Elaboración propia con datos del banco.

La **Figura 3** muestra que este cliente posee una alta tasa de apertura (mayor a 75% generalmente) frente a una gran cantidad de correos que el banco le envió algunos meses del año 2021. Es importante notar que entre marzo y junio del mismo año, esta persona abrió al menos 10 mails mensuales. Por ende, no existiría razón para que en julio del 2021 se le envíen 5 correos, siendo que el cliente demostró que es capaz de leer una cantidad mayor de emails.

El gráfico de la **Figura 4** refleja la apertura de correos por parte de un cliente que se caracteriza por leer pocos emails. Más aún, entre marzo y julio del año 2021, a esta persona se le enviaron menos de 10 correos mensualmente, de los cuales leyó un par entre marzo y abril. Frente a esto, no se justifica que el cliente reciba cerca de 80 mails entre agosto y octubre del mismo año, si su historial deja entrever que suele no leer los correos que recibe por parte del banco.

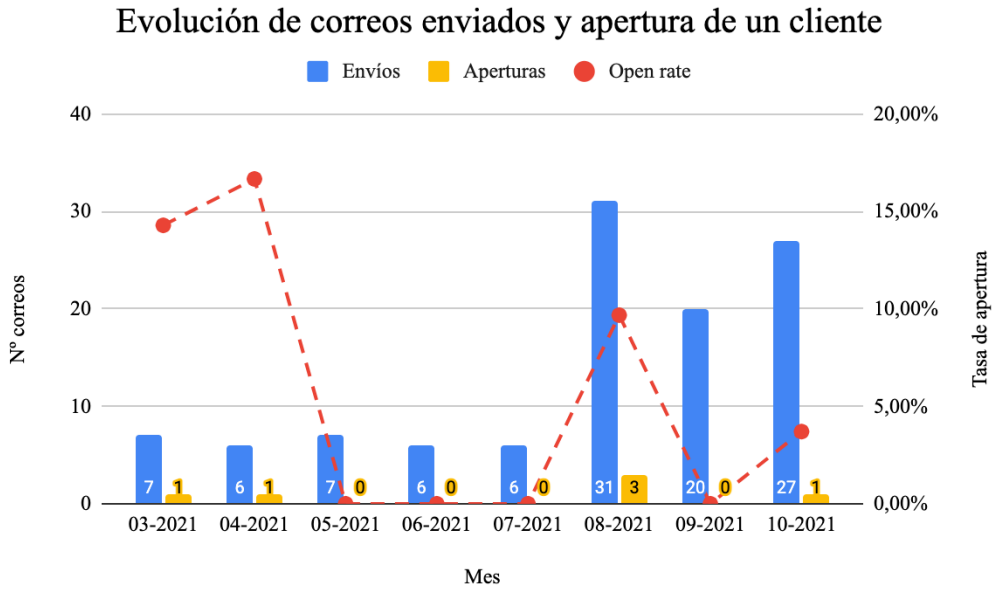

Figura 4: Cantidad de correos enviados y leídos por un cliente con baja tasa de apertura.
Fuente: Elaboración propia con datos del banco.

2.2.2 Desuscripción de clientes

La desuscripción de clientes, en el caso del email, se atribuye a que las políticas de envío de correos no toman en cuenta el grado de saturación[8] de las personas. Según el registro de clientes que el banco no puede contactar, entre el segundo semestre del año 2020 y la primera mitad del año 2021, cerca de 12 mil personas solicitaron dejar de ser contactadas a través de correo electrónico. En otras palabras, en aproximadamente un año, se desuscribieron entre 1 y 2 personas por hora del email marketing del banco, lo que evidencia una constante disminución del alcance que podría tener este *emailing*.

Además, un análisis realizado con información de algunos productos muestra el efecto negativo que tiene el hecho de que los clientes se desuscriban, sobre la venta de estos mismos artículos. En el caso de créditos de consumo, el monto total vendido a personas que registraron su desuscripción entre noviembre del año 2019 y octubre del 2021, asciende a más de $51,7 mil millones de pesos para los clientes que adquirieron un crédito de estos antes de desuscribirse, mientras que las ventas rondan los $33 mil millones de pesos cuando los clientes contrataron el producto luego de registrar su desuscripción.

Esta disminución también se evidencia para créditos hipotecarios y seguros de automóvil, los cuales, bajo el mismo análisis, muestran que el producto hipotecario disminuye sus ventas desde aproximadamente $63 mil millones de pesos antes de que los clientes se desuscriban, hasta un monto cercano a $58 mil millones de pesos luego de que las personas se desuscribieron. Mientras

[8] Se entiende como una persona saturada aquella que se encuentra desuscrita del email marketing del banco o que ha leído pocos de los correos que se le han enviado durante un periodo de tiempo.

que para el seguro, bajan desde aproximadamente $50,4 mil UF hasta $44,9 mil UF, respectivamente.

En resumen, considerando la demanda de productos por parte de personas desuscritas, este análisis evidencia que las ventas disminuyen en 36,12% para el caso de créditos de consumo, 8,05% para créditos hipotecarios y 10,94% para seguros de automóviles, si se compara la adquisición de estos productos antes y después de que los clientes registren sus desuscripciones.

3. Objetivos

3.1 Objetivo general

El objetivo general de este trabajo de título es establecer una política de toques diferenciada por segmentos de clientes en el email marketing del CRM Personas del banco, para optimizar la apertura y la desuscripción de las campañas de *emailing* de la organización.

3.2 Objetivos específicos

- Conocer y comprender la política de email marketing que implementa la organización.
- Establecer un criterio de saturación y evaluar la saturación de las personas contactadas por email.
- Identificar variables relevantes de clientes y realizar una segmentación.
- Generar políticas de toques en email según segmentos de clientes.
- Diseñar experimentos para evaluar la respuesta de las personas frente a diferentes políticas de toques.

4. Marco conceptual

En esta sección se detallan los modelos matemáticos utilizados a lo largo del trabajo. Además, se definen los criterios que se usan para comparar estos modelos y el método de muestreo considerado para determinar los conjuntos de datos incorporados en la modelación.

4.1 Segmentación

Para efectos de este trabajo, segmentación (o *clustering*) se entiende como el proceso en que se divide una base de clientes en grupos (segmentos o *clusters*) compuestos por individuos con características y cualidades similares. En email marketing, esta técnica se puede ocupar para determinar diferentes estrategias de *emailing* para contactar a distintos grupos.

En esta memoria se utilizan tres modelos de este estilo: Agrupamiento jerárquico aglomerativo, *K*-medias y *K*-vecinos más cercanos.

1. **Agrupamiento jerárquico aglomerativo[1][13]**: Tipo de agrupamiento en que se establece un orden jerárquico inicial que cuenta con un segmento por observación. A medida que se avanza en la jerarquía, estos grupos se combinan hasta formar, de no existir alguna restricción, un único segmento mayor que posee todos los registros. Para llevar a cabo este *clustering*, se debe indicar una medida de disimilitud de grupos, la cual se compone por una métrica y un criterio de enlace:

    a. **Métrica**: Corresponde a una medida de la cercanía entre dos observaciones *a* y *b*. Para efectos de este trabajo, se utiliza como métrica la "distancia euclidiana", la cual se define a continuación:

    $$\left|\left|a-b\right|\right|_2 = \sqrt{\sum_i (a_i - b_i)^2} \quad (1)$$

    b. **Criterio de enlace**: Criterio que establece la distancia entre grupos de observaciones *A* y *B*, en función de las distancias entre pares de observaciones (*a,b*). Para efectos de esta memoria, se usa como criterio el "enlace completo", el cual utiliza la distancia máxima entre los pares de observaciones (*a,b*). Su expresión se puede ver en la siguiente fórmula:

    $$max\,\{d(a,b): a \in A, b \in B\} \quad (2)$$

2. ***K*-medias[14]**: Método de agrupación que divide un conjunto de *n* observaciones en *k* grupos, los cuales poseen un valor medio o "centroide", que sirve para asignar las observaciones en estos *clusters*. En esa línea, los datos de cada grupo se caracterizan por ser los más cercanos a este centroide.

    Dado un conjunto de observaciones $(x_1, x_2, \dots, x_n)$, *K*-medias divide estos datos en *k* grupos S_k que minimizan la suma de las distancias cuadráticas entre las observaciones y el centroide de su mismo grupo. Nuevamente, la distancia utilizada en este caso es la euclidiana. De esta forma, la formulación de *K*-medias es la siguiente (con μ_i el valor medio de S_i):

    $$argmin_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \left|\left|x_j - \mu_i\right|\right|_2^2 \quad (3)$$

    Dado un conjunto inicial de k centroides arbitrarios $m_1^{(1)}, m_2^{(1)}, \dots, m_k^{(1)}$; la metodología *K*-medias itera entre dos pasos: el paso de asignación y el paso de actualización.

a. **Paso de asignación**: Sirve para asignar cada observación al grupo que tenga el centroide más cercano. Su expresión matemática se puede ver a continuación:

$$S_i^{(t)} = \{x_p: \left\| x_p - m_i^{(t)} \right\|_2 \leq \left\| x_p - m_j^{(t)} \right\|_2 \ \forall\, 1 \leq j \leq k\} \ (4)$$

b. **Paso de actualización**: Se utiliza para calcular los nuevos centroides luego de las asignaciones. Su formulación queda de la siguiente manera:

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \ (5)$$

Finalmente, cuando las asignaciones de observaciones en cada grupo ya no cambian, se considera que el algoritmo converge y forma los *k* grupos.

3. ***K*-vecinos más cercanos (*K*-nn)[15]**: Método de clasificación en que se estima la función de densidad de probabilidad (o simplemente la probabilidad) de que una observación pertenezca a un determinado *cluster*, a partir de un conjunto de datos de entrenamiento. Así, *k*-nn utiliza las *k* observaciones más cercanas a una nuevo dato, para determinar el segmento que más vecinos tiene de esta nueva observación.

Dado un nuevo registro x_q, que debe ser clasificado según x_1, x_2, …, x_k, sus *k* vecinos más cercanos, se estima una función de densidad de la siguiente forma:

$$F(x_q) \leftarrow argmax_{v \in V} \sum_{i=1}^{k} [v = f(x_i)] \ (6)$$

Donde *V* corresponde al conjunto de segmentos y $f(x_i)$ corresponde al segmento que pertenece el valor x_i . De esta forma, el estimador resultante $F(x_q)$, corresponde al *cluster* que más se repite para los vecinos cercanos de x_q . Por ejemplo, si k = 1, el vecino más cercano a x_q determina $f(x_q)$, su segmento.

4.2 Predicciones

Para el diseño de las políticas de toques, se utilizan Regresiones y Árboles de decisión. Estos modelos matemáticos permiten predecir la cantidad de envíos y los horarios de envío que más se adecuen a las personas, según sus niveles de lectura.

4.2.1 Cantidad de correos a enviar

Para determinar el número de mails a enviar a los clientes, se utilizan modelos de Regresión lineal múltiple y Árbol de regresión.

1. **Regresión lineal múltiple[24]**: Tipo de regresión que supone una correlación entre un conjunto de variables con el valor de otra variable. En otras palabras, este modelo permite medir el efecto de múltiples variables independientes sobre una variable dependiente. Su formulación se presenta a continuación:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \quad (7)$$

   Con:

   - y_i : Variable dependiente para la observación *i*.
   - $x_{k,i}$: Variable independiente *k* para la observación *i.*
   - β_k : Estimador del efecto de la variable independiente *k*, sobre la variable dependiente.

2. **Árbol de regresión[10][18][20]**: Los Árboles de decisión son técnicas predictivas que consisten en una división jerárquica y secuencial de un evento, donde cada división (también conocidas como nodos) describe posibles decisiones y sucesos. De esta forma, es posible evaluar los sucesos resultantes al combinar diferentes decisiones, con lo que estos Árboles permiten examinar gráficamente los resultados y verificar cómo fluye el modelo.

   El procedimiento de un Árbol de decisión crea un modelo de clasificación y pronostica valores de una variable dependiente, a partir de los valores de un conjunto de variables independientes. Entre los algoritmos más conocidos para realizar estas clasificaciones se encuentran: ID3, C4.5, C5.0 y CART, siendo este último el utilizado en este trabajo.

   CART (*Classification and Regression Trees*) es un algoritmo binario que realiza particiones de datos utilizando el "índice de Gini", el cual se usa para calcular la desigualdad de ingresos entre los ciudadanos de un territorio. Aún así, este índice sirve para medir cualquier forma de distribución desigual[3].

   De esta forma, los Árboles de regresión son un tipo de Árbol de decisión CART que se caracterizan porque su variable dependiente es continua.

Para los modelos explicados en esta sección, la variable dependiente corresponde a la cantidad de correos que las personas abren mensualmente en promedio. Mientras que, entre las variables

independientes, se encuentran algunas categóricas como el género de la persona o su nivel educacional; y otras numéricas como la edad de la persona o su antigüedad como cliente del banco.

4.2.2 Horarios de envío

Para encontrar los días y las horas a enviar los correos, se utilizan los modelos de Regresión logística binaria y Árbol de clasificación.

1. **Regresión logística binaria[32]**: Esta regresión se utiliza para modelar la probabilidad de ocurrencia de un evento en función de otros. De esta forma, sirve para predecir el resultado de una variable dependiente dicotómica, en función de un conjunto de variables independientes. Su formulación es la siguiente:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \quad (8)$$

Con:

- p_i: Probabilidad de ocurrencia de un evento para la observación *i.*
- $x_{k,i}$: Variable explicativa *k* para la observación *i.*
- β_k: Estimador del efecto de las variable explicativa *k*, sobre el *odd ratio*[9].

2. **Árbol de clasificación[10][18][20]**: Análogo a lo explicado en la sección anterior para el modelo de Árbol de regresión, los Árboles de clasificación son el otro tipo de Árbol de decisión CART, los cuales se reconocen porque su variable dependiente es categórica.

Para los modelos expuestos en esta parte, la variable dependiente corresponde a la apertura o lectura de los correos que envía la organización. Es decir, el evento tiene como opciones: el cliente abre el correo o el cliente no abre el correo. Por otro lado, entre las variables explicativas se encuentran factores que pueden alterar la probabilidad de apertura de un mail, tales como la clase del mail que se envía[10]; o variables temporales (hora de envío y días de envío).

Es importante mencionar que para los modelos de Árboles de decisión es necesario definir un par de parámetros: el número mínimo de observaciones que puede tener un nodo, y el número mínimo de datos que debe que tener un nodo para ser divisible. Para efectos de este trabajo, ambos parámetros se consideran lo menos exigentes posibles para obtener una mayor cantidad de divisiones por Árbol. De esta forma, toman los valores de 1 y 2 respectivamente.

[9] Corresponde a la razón entre la probabilidad de ocurrencia de un evento y la probabilidad de no ocurrencia del mismo.
[10] La clase de un correo se determina con una combinación entre el ámbito y el tipo de mail. Para más detalle, véase la sección "Desarrollo metodológico: Preparación de los datos – Preparación bases de datos".

4.3 Muestreo aleatorio simple

Algunas veces se hace inviable incorporar todas las observaciones de una base de datos en un modelo matemático. Por esto, se hace necesario contar una muestra representativa de esta población de datos, la cual se puede obtener a partir de una técnica de muestreo.

El muestreo es un proceso que permite obtener una muestra de datos de una población, mediante metodologías probabilísticas, las cuales entregan la probabilidad que tiene cada dato de ser parte de la muestra; y metodologías no probabilísticas, donde las observaciones de una muestra se eligen de acuerdo a ciertos criterios que el investigador considera al momento de realizar el muestreo[27].

En este trabajo se utiliza la técnica de muestreo aleatorio simple, tipo de metodología probabilística que se caracteriza por garantizar que todos los datos de la población tienen la misma probabilidad de ser parte de una muestra. Así mismo, las muestras de un mismo tamaño son igualmente probables de ser obtenidas[29].

Generalmente, este método obtiene las muestras asignando un número único a cada observación de la población, para posteriormente seleccionar aleatoriamente un set de estos mismos números, según el tamaño de la muestra buscada[33]. Cabe destacar que es importante que la muestra resultante sea representativa de la población, lo cual se puede lograr cuidando las distribuciones de algunas variables presentes en los datos.

4.4 Criterios de evaluación

A continuación, se describen ciertos parámetros y métricas que se utilizan para comparar los resultados que se obtienen en el desarrollo de este trabajo.

1. **Método del codo[35]**: Técnica que se utiliza para determinar el número óptimo de *clusters* a considerar en una segmentación. Este método consta de un gráfico que en su eje horizontal posee el número de *clusters* evaluados, mientras que en el eje vertical presenta la distancia intracluster, es decir, la distancia media entre las observaciones de cada *cluster* y sus respectivos centroides.

   Se espera que mediante este método, el grafico resultante presente un punto de inflexión en su curva, al cual se denomina como “codo”. Este codo permite identificar la cantidad óptima de *clusters k* a utilizar en una segmentación, ya que la inflexión deja entrever que un incremento de *k*, no presenta mayor mejora en la distancia intracluster.

2. **Método de la silueta[35]**: Criterio alternativo al descrito anteriormente que permite determinar el número óptimo de segmentos. Esta técnica muestra un gráfico que en el eje *x* posee el número

de *clusters* evaluados, y en el eje *y* muestra el coeficiente de silueta promedio, que es un reflejo de la distancia de separación entre segmentos.

Un coeficiente de silueta indica qué tan cerca está una observación de un *cluster*, respecto a datos de segmentos vecinos. Este parámetro es un número continuo que varía en el rango [-1, 1], donde el valor -1 evidencia que una observación podría estar asignada a un segmento que no le corresponde, mientras que el valor 1 permite concluir que un dato está en el segmento correcto y alejado de los *clusters* vecinos. Este coeficiente se calcula de la siguiente forma:

$$S_{i,k} = \frac{b_{i,k} - a_k}{max(a_k, b_{i,k})} \quad (9)$$

Con:

- $S_{i,k}$ = Coeficiente de silueta para la observación *i* del segmento *k.*
- a_k = Distancia media intercluster del segmento *k.*
- $b_{i,k}$ = Distancia media entre la observación *i* del segmento *k*, y las observaciones de *clusters* vecinos.

De esta forma, el número óptimo de segmentos es aquel que mediante este método, maximice el coeficiente de silueta promedio para el rango de *clusters* evaluados.

3. ***R*²[21]**: Medida estadística de ajuste que refleja la porción de varianza de la variable dependiente que es explicada por las variables independientes. Otros significados de este parámetro es indicar qué tanto se ajusta un modelo a los datos, o qué tanto sirve para realizar predicciones de la variable dependiente. Dada esta última interpretación, que es la que se utiliza mayormente en este trabajo, se presenta la fórmula para calcular R^2:

$$R^2 = 1 - \frac{SSR}{SST} \quad (10)$$

Con SSR la suma de cuadrados residuales, que se calcula como la sumatoria de diferencias cuadráticas entre el valor real de la variable dependiente y su valor predicho:

$$SSR = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \quad (11)$$

Y SST la suma total de cuadrados, que se obtiene con la sumatoria de diferencias cuadráticas entre el valor real de la variable dependiente y su valor promedio:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y_i})^2 \quad (12)$$

De esta forma, R^2 varía en el rango [0,1], entendiéndose que si un modelo se ajusta mejor, R^2 es cercano a 1, mientras que se acerca a 0 cuando el modelo tiene un peor desempeño.

4. **MAE, MAPE y RMSE[37]**: Estas métricas permiten medir la precisión de una predicción. El MAE (*Mean Absolute Error* en inglés), se obtiene a través del promedio de las diferencias absolutas entre la variable dependiente real y su predicción. En el caso del MAPE (*Mean Absolute Percentage Error* en inglés), los errores absolutos individuales son divididos por la observación real. Mientras que el RMSE (*Root Mean Squared* Error en inglés), se obtiene calculando la raíz cuadrada de la media cuadrática de las diferencias entre el dato real y su predicción. La formulación de cada uno se presenta a continuación:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n} \quad (13)$$

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{y_i}}{n} \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n}} \quad (15)$$

Dado que las métricas MAE, MAPE y RMSE están asociadas a mediciones de error en un pronóstico, se entiende que mientras más cercanos a 0 sean estos parámetros, por lo general, el modelo predictor tiene un mejor desempeño.

5. ***Residual deviance***: La desviación residual es un parámetro que permite concluir sobre el ajuste de un modelo de predicción. Se obtiene mediante una diferencia entre la verosimilitud del modelo saturado, que es el que cuenta con toda la varianza de los datos, por lo que se ajusta perfectamente; y la del modelo propuesto, que es el que efectivamente realiza un pronóstico[23]. Su fórmula general se presenta a continuación[8]:

$$D(y, \hat{y}) = 2(log(p(y|\widehat{\theta_s})) - log(p(y|\widehat{\theta_p}))) \quad (16)$$

Con:

- y: Datos observados o reales.
- $\hat{y}$: Datos predichos.
- $\widehat{\theta_s}$: Parámetros estimados por el modelo saturado.
- $\widehat{\theta_p}$: Parámetros estimados por el modelo propuesto.

De esta forma, mientras más baja sea la desviación de residuos, se entiende que el modelo y el pronóstico tienen mejor desempeño.

5. Metodología

La metodología que se utiliza para desarrollar este trabajo es CRISP-DM (*Cross Industry Standard Process for Data Mining*, 1996), que corresponde a un proceso que permite describir el ciclo de vida de un proyecto de análisis de datos. Las etapas que incluye, algunas bidireccionales, son: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelación, Evaluación y Despliegue[4]. En la siguiente figura es posible ver el esquema de esta metodología:

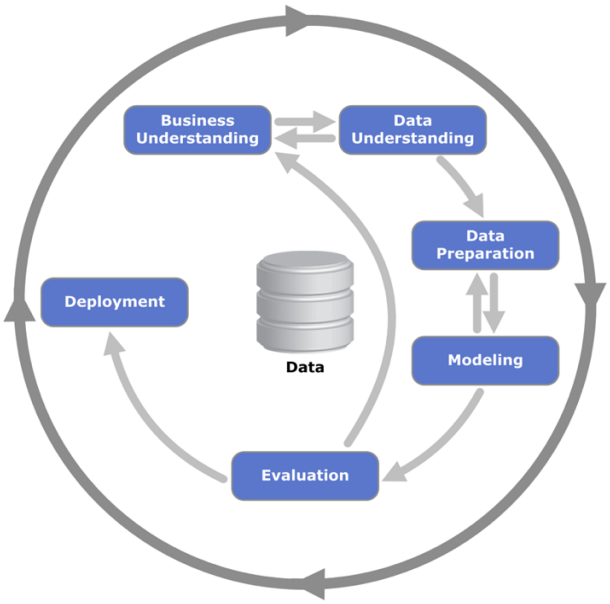

Figura 5: Metodología CRISP-DM.
Fuente: CRISP-DM: Una metodología para minería de datos, Health Data Miner.

5.1 Comprensión del negocio

En esta etapa se analiza la situación actual del email marketing dentro del banco. En ese sentido, se reconocen los procesos con los que se realiza el *emailing* en el CRM Personas, y las principales variables que influyen en las políticas de envío de correos utilizadas por el área, donde se identifican ciertas restricciones que son consideradas a la hora de mandar los emails. Además, aquí se configuran los permisos para acceder a las bases de datos con las que se desarrolla este estudio, las que son tratadas a través del programa *Teradata SQL Assistant*.

5.2 Comprensión de los datos

Se estudian las bases de datos facilitadas por la empresa, para así entender la información con la cual se está realizando esta memoria. La principal base con que se trabaja, contiene el registro de los correos que han sido enviados por la organización desde noviembre del 2019, donde se verifica que la información es agregada diariamente, por lo que la base es actualizada constantemente. Esta contiene información de la persona a la que se le envía el mail, datos del correo y observaciones sobre el envío mismo.

La base descrita anteriormente servirá fundamentalmente para analizar la cantidad de correos que abren o leen las personas. Por ende, es necesario identificar cuándo un correo efectivamente es leído. Ante esto, aparecen registros que debiesen ser descartados o corregidos, tales como correos que son abiertos pero no para ser leídos, datos erróneos o mails que no tienen como finalidad comunicar a un cliente.

Una segunda base a tener en consideración en este trabajo posee información sobre los clientes que se han desuscrito de las campañas de marketing del banco, y por ende, no pueden ser contactados nuevamente. Esta contiene registros desde el año 2009, y sirve en este estudio para identificar aquellas personas que en particular solicitaron no recibir más correos de la empresa.

5.3 Preparación de los datos

Previo a la etapa de preparación de los datos, se analiza el grado de saturación de los individuos contactados por email. Para esto, se selecciona un criterio de saturación que permite identificar a aquellas personas que en un periodo de tiempo se han caracterizado por leer pocos correos.

Una vez elegido el criterio de saturación, se confeccionan dos bases de datos que sirven para desarrollar los modelos de la etapa siguiente. La primera, denominada “Base de personas”, contiene información sobre los clientes que han sido contactados por email entre julio del año 2020 y octubre del año 2021. Entre las variables incorporadas en esta base se encuentran datos para identificar a las personas, un par relacionadas a los productos que posee cada cliente, y otras que se obtienen según los niveles de lectura de correos.

La segunda base, nombrada “Base de correos”, posee el registro de emails que se enviaron entre octubre del año 2020 y octubre del 2021, a los individuos que componen la Base de personas. Así, la Base de correos posee información sobre la fecha de envío de los mails, datos relacionados al contenido del correo y un indicador para reconocer si el email fue leído o no.

Finalmente, mediante el criterio de saturación seleccionado, se establecen ciertas condiciones que deben cumplir los niveles de lectura de los clientes, para que estos sean considerados como

saturados. Con esto, es posible incorporar una variable dicotómica en la Base de personas para reconocer si un individuo está saturado o no.

5.4 Modelación y evaluación

En este estudio, las etapas de modelación y evaluación que plantea la metodología CRISP-DM, se trabajan paralelamente. En otras palabras, se comparan los resultados que entregan los modelos matemáticos y el desempeño de estos últimos, para con esto, determinar las técnicas que se implementan en la etapa de despliegue.

5.4.1 Segmentación

En esta parte se realiza un *clustering* mediante los métodos de Agrupamiento jerárquico aglomerativo y *K*-medias, a varias muestras de datos pertenecientes a la Base de personas, las cuales son obtenidas a través de Muestreo aleatorio simple. Estas segmentaciones permiten conocer las variables a incorporar en los modelos y el número de *clusters* que se evalúan para considerar en la segmentación.

Posteriormente, se revisa el flujo que siguen los modelos para realizar el *clustering*, según diferentes números de grupos a formar. Esto, en conjunto con el Método de la silueta, permite escoger la técnica, la muestra y los segmentos que precisan los datos de entrenamiento que se utilizan en *K*-vecinos más cercanos, para así clasificar a todos los individuos registrados en la Base de personas en los *clusters* definidos.

5.4.2 Cantidad de correos a enviar

Con las técnicas de Regresión lineal múltiple y Árbol de regresión, se predice un intervalo del número de correos a mandar por persona. Estos pronósticos se realizan, en un principio, sin especificar el segmento al que pertenecen los individuos. Sin embargo, el ajuste y desempeño de los modelos exige realizar las predicciones según el *clustering* obtenido, de forma tal que se determinan rangos del número de mails a enviar a las personas, según el segmento al que estas pertenezcan.

A pesar de incorporar la segmentación en estos pronósticos, los resultados se caracterizan por sus altos niveles de error. No obstante, la predicción del intervalo de envíos a utilizar con un *cluster* en particular, destaca por presentar mejores métricas de desempeño.

Así, según la modelación y los criterios de evaluación que se utilizan en esta parte, es posible seleccionar una técnica matemática que, con una nueva interpretación de los resultados, define la cantidad de correos a enviar por persona.

5.4.3 Horarios de envío

En esta parte de la modelación se trabaja con la Base de correos, la cual es particionada según la clase de un correo. De esta forma, se obtienen 20 clases para las cuales se predice, mediante las técnicas de Regresión logística binaria y Árbol de clasificación, una probabilidad de apertura de los mails según el horario en que se envía.

Los resultados obtenidos en esta parte, en conjunto con el criterio de evaluación, dan paso a una discusión sobre la técnica a utilizar para seleccionar el día y la hora de envío de los correos. Además, se verifica que existen casos en que un horario que predice la probabilidad de apertura más alta, pareciera no ser el mejor ya que existen combinaciones de día y hora que predicen una mejor posibilidad de apertura. Sin embargo, este último par se caracteriza por ser un horario que el banco no suele utilizar para enviar los respectivos correos.

Finalmente, se escoge un modelo matemático y se incorporan los segmentos de personas en la modelación, para así definir los horarios a utilizar en el envío de los diferentes correos.

5.5 Despliegue

En esta última etapa de la metodología, se hace entrega a la organización de los resultados y las principales conclusiones que deja el trabajo. Así, se describen las características que presentan los segmentos de clientes y las formas que parecerían ser más adecuadas para que el CRM Personas lleve a cabo el email marketing de la organización. También, se detallan ciertos horarios de envío de mails que se podrían tener en cuenta en las políticas de toques, los cuales parecerían mejorar las tasas de lectura a pesar de que el banco no los implemente típicamente.

Además, en esta parte se lleva a cabo un diseño experimental, el cual se espera que sirva para probar directamente los resultados de este trabajo en el *emailing* de la organización, y así establecer la forma con que se debiese contactar a los distintos clientes a través de correo electrónico.

6. Alcances

- El trabajo se realiza con bases de datos históricas que posee la organización, es decir, no se llevan a cabo implementaciones para obtener datos.
- Se consideran sólo clientes cuentacorrentistas[11] de tipo Personas.
- Se toma en cuenta sólo el canal de email.
- No se modifican las campañas de email marketing que desarrolla el banco.
- No se realizan cambios a los correos enviados en las campañas de *emailing*.

[11] Personas con una cuenta corriente activa en el banco.

- El grado de cumplimiento de las metas del banco no son un indicador para evaluar los resultados de este trabajo.

7. Desarrollo metodológico

7.1 Comprensión del negocio

Desde una mirada general, el proceso que lleva a cabo el banco para enviar correos comienza con una selección del público a contactar a través de este medio. Este público objetivo es priorizado para establecer aquellas personas a las cuales es factible contactar, según la política de envíos que utiliza el banco. Luego, se definen los grupos de control, que determinan los individuos que efectivamente son contactados a través de email. Posteriormente, se generan los archivos de salida que permiten configurar los correos que son enviados al público objetivo final. Por último, se realizan los envíos de correos y se recibe la respuesta a estos mismos, la cual se almacena y se considera en futuros procesos de envío de correos. La siguiente figura muestra un resumen del proceso de envío de correos:

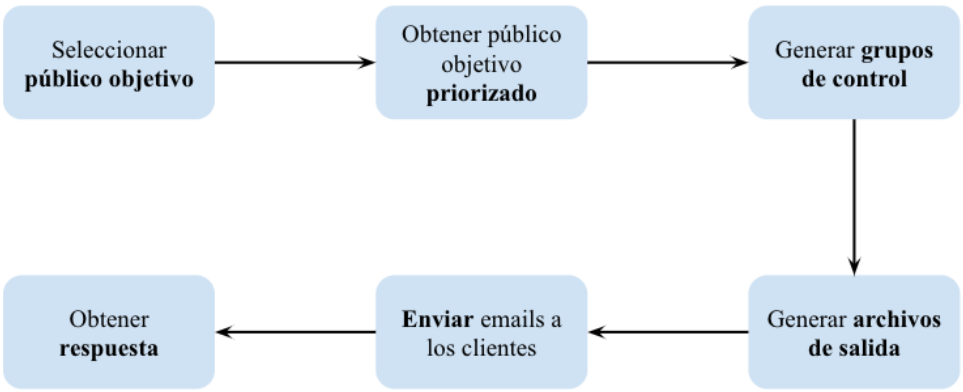

Figura 6: Proceso de envío de correos.
Fuente: Facilitado por el banco.

La política de envíos que suele utilizar el banco, considera las siguientes variables y parámetros relacionados a la política de saturación (Saturación y *Recency*), a la clase del email a enviar (Prioridad estratégica y Obligatoriedad) y a la priorización (*Score*, Origen y Matriz de priorización):

- Saturación: Corresponde al máximo de correos que se le pueden enviar a una persona en un periodo de tiempo. La política de envíos establece, de forma general[12], que se pueden enviar a

[12] Independiente del correo que se envíe y a quien se le envíe.

lo más 1 email diario, máximo 7 en una semana y a lo sumo 15 en un mes. Sin embargo, dependiendo de la información que contengan los correos, se pueden superar estos límites.

- *Recency*: Define el número mínimo de días a esperar antes de volver a contactar a una persona por mail. Suele ser 2 o 3 días.
- Prioridad Estratégica: Permite otorgar diferentes prioridades a los correos por enviar. De esta forma, se definen los correos que no se deben enviar para cumplir con la política de saturación.
- Obligatoriedad: Define los correos que deben enviarse a pesar de lo que señale la política de saturación.
- *Score*: Corresponde a un puntaje que se asigna a los clientes, el cual se relaciona con la información contenida en el correo. A mayor puntaje, se espera que mayor sea la receptividad del mail por parte del cliente.
- Origen: Define una prioridad de envío de un correo según el motivo del *lead.* Puede ser "Spot", que son mails que se envían buscando aprovechar una oportunidad; o "Recurrente", que son correos que se envían para no perder el contacto con la persona.
- Matriz de Priorización: Matriz que se ocupa para desempatar paridades que pueden existir al definir una prioridad de envío de un correo mediante el *Score* y el Origen.

Este tipo de regulaciones internas que establece el banco, buscan que las campañas de email marketing no sean percibidas como *spam*, para así no saturar al cliente ni aumentar los niveles de desuscripción. En este caso, la organización no debiese incluir en su *emailing* a aquellos clientes que se desuscriben, ya que estos últimos, amparados por la Ley Nº 19.496 (conocida como Ley del Consumidor), están en su derecho de solicitar la suspensión del envío de correos que no deseen. Así, el artículo 28-B de la Ley del Consumidor, prohíbe el envío de email marketing a aquellos clientes que exigen su remoción de las bases de destinatarios[17].

Otro aspecto a considerar es que, tal como se menciona en la sección "Introducción: Email Marketing", el costo marginal de enviar un correo adicional, en las campañas de *emailing* del banco, es despreciable. Esto tendería a fomentar el envío masivo de correos, situación que no ocurre en el área de CRM Personas. Hoy en día, los proveedores de correo electrónico (*Mail Service Provider* en inglés) tales como *Gmail*, *Outlook* o *iCloud Mail,* entre otros, poseen herramientas que permiten clasificar como *spam* aquellos correos masivos que incrementan drásticamente su cantidad de envíos[5]. De esta forma, el número de correos enviados en las campañas de email marketing del área CRM Personas, debiese aumentar paulatinamente para no caer en la categoría de *spam*.

7.2 Comprensión de los datos

La principal fuente de datos que se utiliza en este trabajo corresponde a una base que contiene el historial de correos que han sido enviados desde noviembre del 2019. Desde esa fecha hasta octubre del año 2021, la base registra más de 215 millones de filas[13], sin embargo, todos los días se le agrega información, por lo que la cantidad de datos aumenta constantemente. Denominada para efectos de este trabajo como "Base de envíos", permite calcular la cantidad de envíos, aperturas y

[13] Una fila corresponde a un correo que fue enviado.

tasas de lectura de las personas en algún intervalo de tiempo. La figura a continuación muestra la evolución de la cantidad de correos que se han sido enviados por el banco mes a mes:

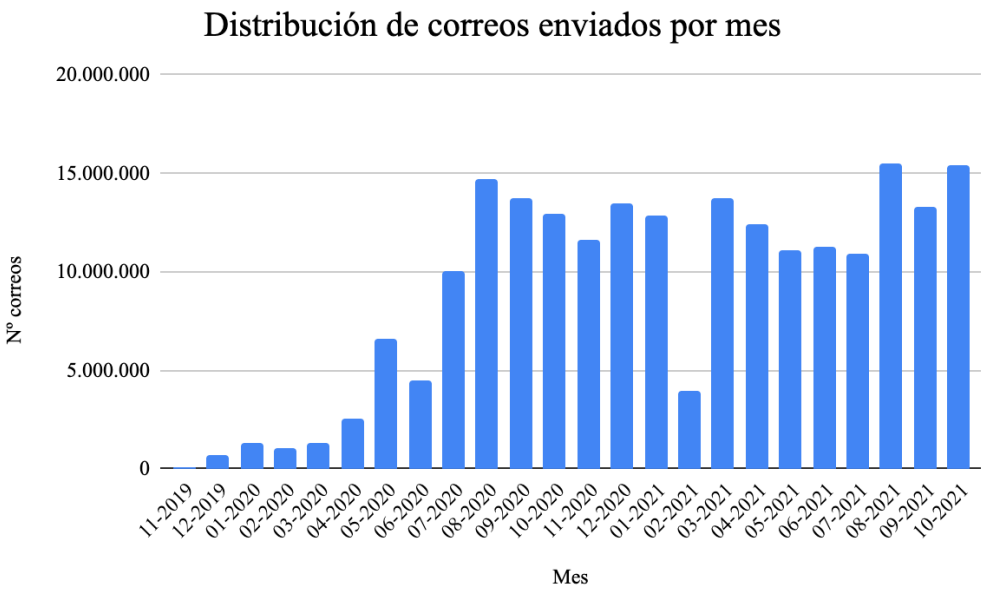

Figura 7: Cantidad de correos enviados mensualmente.
Fuente: Elaboración propia con datos de la Base de envíos.

El gráfico anterior muestra que el registro de los correos enviados por el banco se consolidó entre julio y agosto del año 2020, ya que a partir de esa fecha, la cantidad de datos se estabilizó por sobre los 10 millones de registros mensuales. Esto se explica principalmente por la regulación del registro de información en la Base de envíos, la cual para ese periodo comenzó a registrarse diariamente.

Por otro lado, destaca en la **Figura 7** la columna referente a febrero del 2021, mes que posee cerca de 4 millones de datos y que se aleja bastante de la cantidad media de información que poseen los meses del semestre previo y posterior. Si bien no existe mayor evidencia que explique este hecho, se cree que esta disminución de información se debe a que febrero típicamente es un mes caracterizado por las vacaciones que toma el personal de la organización y los clientes, por lo que baja la cantidad de contactos entre estos últimos y el banco.

La base en cuestión, cuenta con variables relacionadas al cliente como su RUT o su dirección de correos electrónico, otras con datos de los emails enviados tal como el asunto o la fecha de envío, y variables que hacen referencia a las acciones del cliente sobre el correo, tales como la fecha de apertura o el enlace sobre el que hizo click. A continuación, se presentan las principales variables de esta base según su tipo, de las cuales la mayoría se tienen en consideración en este trabajo:

Tabla 1: Variables presentes en Base de envíos.
Fuente: Elaboración propia.

Variables tipo *"date"*[14]	**Variables tipo *"string"***	**Variables tipo *"integer"***
Fecha de envío	Correo electrónico del cliente	RUT del cliente
Fecha de apertura	Enlace clickeado	
Fecha de click	Asunto del mail	
Fecha de rebote	Nombre del mail	
Fecha programada de envío		

Además, en el siguiente detalle se adjunta una descripción de cada una de las variables:

- Fecha de envío: Posee información sobre el momento en que fue enviado un correo por parte del banco.
- Fecha de apertura: Cuenta con los datos del horario en que un correo fue abierto o leído por una persona.
- Fecha de click: Entrega el horario en que se realizó un click sobre el enlace adjunto en un correo.
- Fecha de rebote: Aparece cuando un correo enviado no logra ser entregado a su destinatario.
- Fecha programada de envío: Indica el momento en que está planificado el envío de un correo.
- Correo electrónico del cliente: Dirección de email al que se envía un correo a una determinada persona.
- Enlace clikeado: Corresponde al enlace adjunto en un correo sobre el que una persona hace click.
- Asunto del mail: Texto referente al asunto indicado en un correo enviado.
- Nombre del mail: Permite determinar la clase del correo enviado.
- RUT del cliente: Número RUT de cada cliente sin el dígito verificador. Sirve para identificar a las personas.

Mediante la variable "Nombre del mail", es posible conocer la clase del correo, más aún, se puede saber el ámbito del correo, lo que se relaciona con el objetivo del mail que se envía. Este posee 6 niveles: Fidelizar, Vender, Informar, Cobrar, Atraer y Pyme. En la tabla a continuación, es posible revisar la proporción de estos ámbitos entre los correos que registra la Base de envíos:

[14] El formato de estas variables es "dd-mm-aaaa hh:mm:ss".

Tabla 2: Proporción de ámbitos en los correos enviados.
Fuente: Elaboración propia con datos de la Base de envíos.

Ámbito	Nº correos [Millones]	Porción de correos
Fidelizar	107,21	49,86%
Vender	39,47	18,36%
Informar	64,70	30,09%
Cobrar	0,62	0,29%
Atraer	2,23	1,04%
Pyme	0,70	0,33%

La tabla anterior muestra que entre los correos registrados en la Base de envíos, casi la mitad de los datos corresponden a envíos que tienen por objetivo fidelizar al cliente, es decir, fortalecer la relación del banco con una persona. En una porción más baja aparecen los correos que buscan que el cliente adquiera un producto y aquellos que se envían para entregar información sobre los procesos de una persona con el banco, es decir, los registros de los ámbitos vender e informar respectivamente.

Por otro lado, aparecen los correos que se envían en una proporción menor tales como: los del ámbito cobrar, que se relacionan netamente con la cobranza de deudas que tenga un cliente con el banco; los del ámbito atraer, que ofrecen planes de cuenta a las personas; y los del ámbito pyme, enfocados en comunicar a emprendedores. También existen casi 94 mil registros de correos a los que no es posible determinar su ámbito, los cuales representan aproximadamente un 0,04% de los datos.

En este punto, es importante describir aquellos registros de correos que en principio se tratarían de manera diferente al resto de datos. Estos se distribuyen en 3 grupos:

- Rebote: Formado por correos que presentan una fecha de rebote.
- Erróneos: Agrupa los registros de mails que pueden presentar 3 condiciones:
    - La fecha de envío es posterior a la fecha de apertura.
    - La fecha de apertura es posterior a la fecha de click.
    - No existe fecha de apertura pero se registra fecha de click.
- Desuscripción: Correos que fueron abiertos para desuscribirse de las campañas de marketing del banco, por lo que no deberían considerarse como leídos. Esto se puede determinar cuando las personas hacen click sobre un enlace de desuscripción contenido en el mail.

En total, estos registros “especiales” suman más de 1,6 millones de observaciones. En el caso de los rebotados y erróneos, estos tipos de datos pueden ser eliminados para los análisis posteriores ya que presentan fallas en su implementación e información respectivamente. En el caso de los correos del grupo desuscripción, la opción es considerarlos como mails que efectivamente fueron enviados, pero que no fueron leídos ya que fueron abiertos para otro fin. En la gráfica siguiente se muestra la evolución de esta clase de correos:

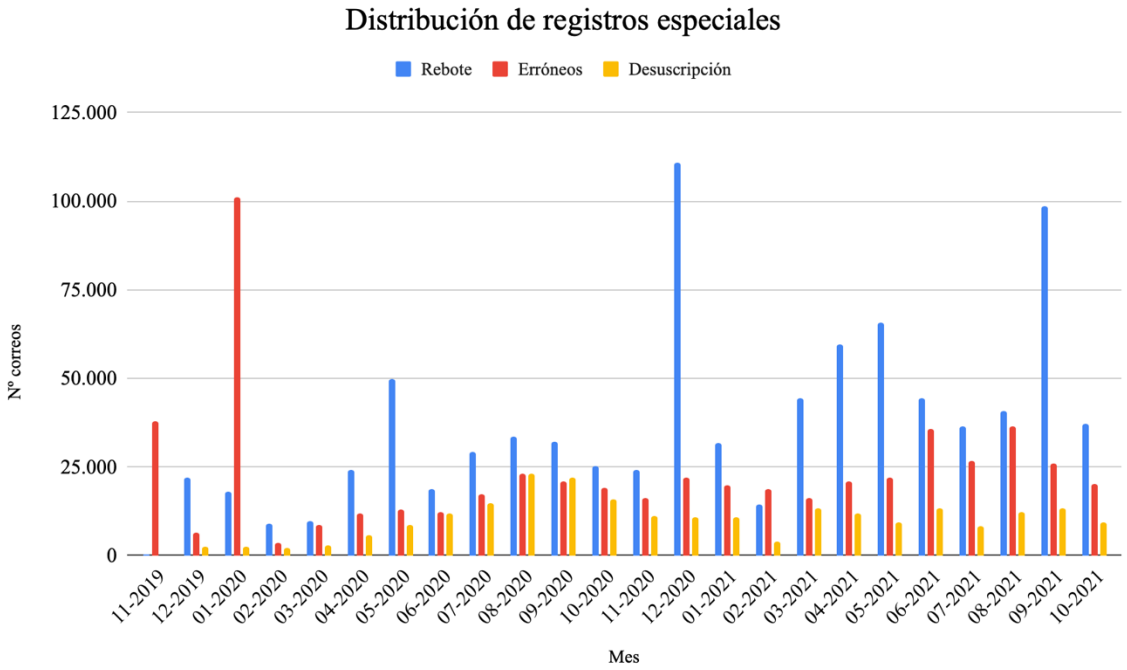

Figura 8: Cantidad de observaciones especiales registradas por mes.
Fuente: Elaboración propia con datos de la Base de envíos.

La figura anterior muestra una clara predominancia del grupo rebote entre los registros especiales, dado que este conjunto, por lo general, es el que más mails agrupa mes a mes. Así es como los correos que no logran llegar a su destinatario suman en total más de 881 mil datos y representan el 52,54% de registros especiales. Más atrás aparecen los conjuntos erróneos y desuscripción, donde se agrupan 556.609 y 239.506 emails respectivamente, significando un 33,18% y 14,28% de los registros especiales correspondientemente.

Finalmente, la otra base de datos que se tiene en consideración a lo largo de este trabajo, es aquella que contiene información sobre las personas que se han desuscrito de las campañas de marketing de la organización. Es decir, esta base posee la lista de individuos a las que el banco no puede volver a contactar mediante sus diferentes canales de comunicación. Los datos han sido registrados a partir de enero del año 2009, y al igual que la base descrita anteriormente, se le agregan datos constantemente. Para octubre del 2021, la base de desuscripción registraba más de 85 mil personas que no pueden ser contactadas a través de correo electrónico. Además, posee variables como el RUT de la persona, el motivo y el canal por el que no se le puede contactar, y la fecha de desuscripción, entre otras. Para efectos de este trabajo, esta base se denomina “Base de desuscripción”

7.3 Preparación de los datos

En esta etapa se estudia la saturación de los clientes contactados por el banco a través de email, es decir, se analiza la cantidad de correos que las personas han leído durante un periodo de tiempo. Además, se detalla la construcción de las bases de datos que son utilizadas en la etapa de modelación.

Cabe destacar que de aquí en adelante se utilizan los datos registrados desde julio del año 2020, mes en que la información comienza a mostrar cierta estabilidad.

7.3.1 Criterio de saturación[15]

Si bien el banco cuenta con una política de saturación, no existe un criterio para medir el grado de saturación de las personas y determinar si un individuo está saturado por recibir correos. Con esto, sería posible encontrar aquellos clientes que potencialmente podrían desuscribirse del email marketing de la empresa. Por esto, el análisis de esta parte comienza estableciendo 3 criterios arbitrarios de saturación en email, los cuales se detallan a continuación:

1. **Criterio 1**: "Un cliente está saturado si ha leído menos de un X% de los correos que se le han enviado los últimos 6 meses".

   Para estudiar el grado de saturación a través de este criterio, se realiza la matriz de envíos/aperturas que se presenta a continuación:

Tabla 3: Matriz de envíos/aperturas.
Fuente: Elaboración propia con datos de la Base de envíos.

Periodo 01/04/2021 – 30/09/2021								
Envíos/Aperturas	[0,10]	[11,20]	[21,30]	[31,40]	[41,50]	[51,60]	[61,70]	>70
[1,10]	842.604	-	-	-	-	-	-	-
[11,20]	68.633	6.456	-	-	-	-	-	-
[21,30]	54.126	6.207	2.552	-	-	-	-	-
[31,40]	87.425	15.497	7.039	4.890	-	-	-	-
[41,50]	47.042	16.957	8.373	6.245	3.977	-	-	-
[51,60]	29.764	8.834	4.778	3.215	2.796	2.359	-	-
[61,70]	22.600	7.941	4.496	2.878	2.153	1.999	1.623	-
>70	136.652	57.095	33.440	22.519	16.511	13.012	11.099	34.536

[15] Los análisis mostrados en esta parte se realizan con los datos registrados entre julio del año 2020 y septiembre del 2021, siendo este último el mes más reciente que contaba con la completitud de sus datos en ese entonces.

La matriz anterior muestra, en su eje vertical, intervalos del número total de correos que se han enviado a personas contactadas entre abril y septiembre del año 2021, mientras que en el eje horizontal, se aprecian intervalos de la cantidad total de emails que han abierto o leído esas personas en el mismo periodo. Así, por ejemplo, existen 842.604 clientes que en 6 meses recibieron entre 1 y 10 correos, de los cuales abrieron entre 0 y 10.

De esta forma, utilizando el criterio planteado con X = 25 por ejemplo, se puede decir que de los casi 1,6 millones de clientes contactados por email entre abril y septiembre del 2021, existen al menos 200 mil personas saturadas. Esto se obtiene si se suman a los clientes que han recibido al menos 41 emails y han abierto a lo más 10. Esto mismo es replicable con diferentes intervalos de tiempo y distintos valores para X.

2. **Criterio 2**: "Un cliente está saturado si no ha leído ninguno de los correos que se le han enviado los últimos X meses".

   Este criterio considera a aquellas personas que se le han enviado emails todos los meses, los últimos X meses. Bajo esto, la base de envíos muestra que existen más de 100 mil clientes que se consideran saturados cuando X = 6 (es decir, han recibido correos en abril, mayo, junio, julio, agosto y septiembre, todos del año 2021). Mientras que hay casi 260 mil personas que estarían saturadas si X fuese igual a 3 (o sea, han recibido emails en julio, agosto y septiembre del 2021).

3. **Criterio 3**: "Caída en la tasa de apertura".

   Bajo este criterio, se consideran saturados aquellos clientes que evidencian, durante varios meses, un descenso en su tasa de lectura. Un ejemplo de este caso se puede apreciar en el siguiente gráfico, el cual muestra la evolución de la tasa de apertura de un cliente en un intervalo de tiempo de más de un año de datos:

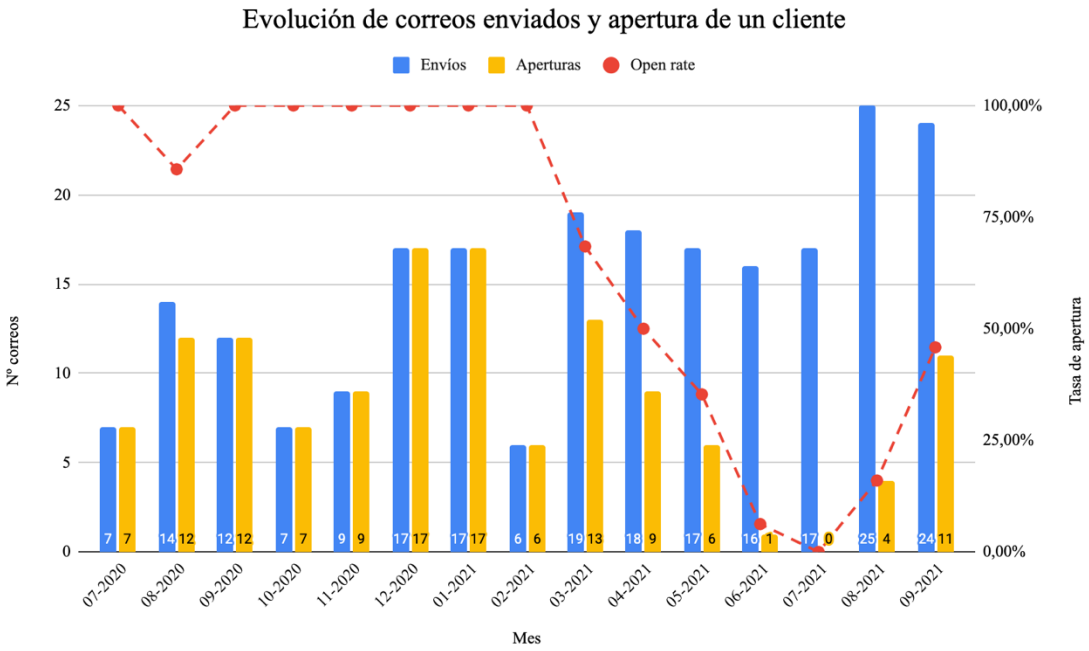

Figura 9: Cantidad de correos enviados y leídos por un cliente con caída en su tasa de apertura.
Fuente: Elaboración propia con datos de la Base de envíos.

En la figura anterior se puede apreciar que entre julio del 2020 y febrero del 2021, el cliente refleja una alta tasa de lectura, más aún, alcanza el 100% de apertura en la mayoría de las ocasiones. Sin embargo, a partir de marzo del 2021, esta tasa comienza a desplomarse, llegando a ser 0% para julio de ese mismo año. Da la impresión de que esta caída es consecuencia de factores que no se relacionan con la cantidad de correos enviados, ya que en un principio, el cliente tuvo una buena tasa de apertura cuando recibía 12, 14 o 17 correos por mes, es decir, evidenció ser capaz de leer varios correos mensualmente. Por ende, se esperaría que el cliente también tuviese un buen *open rate* para la segunda mitad del periodo analizado, contrario a lo mostrado en la **Figura 9**. En conclusión, esta persona podría considerarse saturada dada la disminución que tuvo en sus niveles de lectura mensual.

Finalmente, se decide por trabajar a lo largo de esta memoria con el tercer criterio de saturación, puesto que para la empresa tiene más validez y además, un símil del mismo fue implementado en el trabajo de memoria "Estudio de la Saturación en Email Marketing para un Negocio de Retail*"* de Miguel Gutiérrez, titulado de Ingeniería Civil Industrial en la Universidad de Chile el año 2019.

7.3.2 Preparación bases de datos

En esta sección, se detalla la construcción de un par de bases de datos que sirven para llevar a cabo los análisis presentados en las partes posteriores de este trabajo. Dicho esto, la primera base, denominada como "Base de personas", se confecciona con información de los clientes que el banco les ha enviado al menos un email durante todos los meses comprendidos entre julio del 2020 y

octubre del 2021, sin considerar febrero del 2021 ya que, como se vio en la sección "Desarrollo metodológico: Comprensión de los datos", en promedio tiene menos información respecto al resto de meses en estudio. De esta forma, la Base de personas cuenta con 405.312 filas, donde cada una corresponde al registro de una persona. Esta base contiene datos propios de los individuos, también variables relacionadas a sus niveles de lectura de correos, y otras que contienen información transaccional. En la siguiente tabla se presenta el listado de variables que contiene esta base, según el tipo de variable:

Tabla 4: Variables presentes en Base de personas.
Fuente: Elaboración propia.

Variables tipo *"string"*	**Variables tipo *"integer"***
Género	ID
Estado civil	Edad
Nivel educacional	Antigüedad
Día mayor lectura	Productos
Hora mayor lectura	Nº productos
	Promedio mensual envíos
	Promedio mensual aperturas
	Promedio mensual *open rate*
	Demora promedio de lectura

A continuación, se describe cada una de las variables mostradas en la tabla anterior:

- Género: Señala el sexo de una persona. Posee los siguientes niveles:
  - M: Masculino.
  - F: Femenino.
- Estado civil: Indica el estado civil de un individuo. Posee los siguientes niveles:
  - SEP: Separado/a.
  - CAS: Casado/a.
  - VIU: Viudo/a.
  - CCV: Conviviente.
  - DIV: Divorciado/a.
  - SOL: Soltero/a.
- Nivel educacional: Señala el mayor nivel educacional cursado (o en curso) por una persona. Posee los siguientes niveles:
  - BAS: Educación básica.
  - EUN: Estudiante universitario.
  - MED: Educación media.
  - SIN: Sin educación.

    - TEC: Educación superior técnica.
    - UNV: Educación superior universitaria.
- Día mayor lectura: Abreviación de 3 letras que indica el día que una persona, en total, registró mayor lectura de correos (en inglés). Por ejemplo, si un cliente abrió más correos un día lunes, Día mayor lectura = "Mon". Si no existe un día que supere estrictamente a los demás, la variable toma el valor "Otro".
- Hora mayor lectura: Texto de la forma "Entre hh:00 y hh:59 hrs." que indica el rango de hora en que una persona, en total, registró mayor lectura de correos. Por ejemplo, si un cliente abrió más correos entre las 11:00:00 y 11:59:59 hrs., Hora mayor lectura = "Entre 11:00 y 11:59 hrs.". Si no existe una hora que supere estrictamente a las demás, la variable toma el valor "Otro".
- ID: Número entero que reemplaza al RUT de una persona. Va desde 1 a 405.312.
- Edad: Número entero que indica la edad en años de un individuo.
- Antigüedad: Número continuo que indica los años que una persona ha sido cliente del banco.
- Productos[16]: Indicador binario que toma el valor 1 cuando una persona posee cierto producto, 0 cuando no. Se cuenta con una variable de este tipo por cada artículo listado a continuación:
    - Tarjeta de crédito.
    - Inversiones.
    - Seguros.
    - Cuenta prima.
    - Crédito de consumo.
    - Crédito hipotecario.
- Nº productos[17]: Número entero que señala la cantidad de cierto producto que posee una persona. Cada uno de los artículos listados en la variable anterior, cuenta con una variable de este tipo.
- Promedio mensual envíos: Número continuo que señala la cantidad promedio de correos que recibe una persona por mes. Corresponde al promedio simple entre la cantidad de envíos mensuales.
- Promedio mensual aperturas: Número continuo que señala la cantidad promedio de correos que abre o lee una persona por mes. Corresponde al promedio simple entre la cantidad de aperturas mensuales.
- Promedio mensual *open rate*: Porcentaje que señala la tasa promedio de correos que una persona abre por mes. Corresponde a la división entre Promedio mensual aperturas y Promedio mensual envíos.
- Demora promedio de lectura: Número continuo que indica las horas que una persona, en promedio, tarda el leer un correo.

La Base de personas posee una proporción mayor de clientes hombres al contar con 239.938 individuos de este género. Además, los estados civiles que predominan en los datos son "Solteros/as" y "Casados/as" con 220.242 y 151.639 personas que integran estos niveles, respectivamente. En cuanto a los niveles educacionales, la mayoría de los clientes registra haber completado la educación superior si se tiene en cuenta a los 123.150 profesionales técnicos y 178.299 profesionales universitarios. Cabe mencionar también el rango de edad, que va desde los

[16] Por ejemplo, si una persona posee una tarjeta de crédito y tres inversiones, el indicador para Tarjeta de crédito e Inversiones es 1, mientras que para cada uno de los artículos restantes, el indicador es 0.

[17] Por ejemplo, si una persona posee una tarjeta de crédito y tres inversiones, el nº para Tarjeta de crédito es 1, el nº para Inversiones es 3, mientras que para cada uno de los artículos restantes, el nº es 0.

19 a los 104 años, con una media de 44 años; y el de antigüedad, que varía entre un mes y más de 71 años, con un promedio cercano a los 11 años.

Revisando las variables que hacen referencia a los productos, estas dejan ver que la tarjeta de crédito es el artículo que más clientes abarca, considerando que cerca del 90% de las personas posee al menos una de estas tarjetas. Este resultado puede deberse a que el grupo de clientes analizados son cuentacorrentistas, y este tipo de cuentas generalmente tienen asociadas una tarjeta de crédito. También destaca el hecho de que el seguro es el producto que más se adquiere, ya que en promedio las personas tienen aproximadamente 2 seguros, haciendo que este artículo supere en este aspecto a todo el resto de la cartera. Este dato de los seguros tiene sentido si se considera que las personas pueden tener varios productos de este estilo simultáneamente. Lo descrito anteriormente puede revisarse en mayor profundidad en la tabla a continuación:

Tabla 5: Distribución de productos entre clientes.
Fuente: Elaboración propia con datos de la Base de personas.

Producto	Adquisición	Nº productos promedio
Tarjeta de crédito	89,99%	1,32
Inversiones	25,79%	1,12
Seguros	64,17%	1,67
Cuenta prima	66,49%	0,93
Crédito de consumo	39,64%	0,55
Crédito hipotecario	19,88%	0,21

Dada la tabla anterior, cabe mencionar también que el crédito hipotecario es el producto menos frecuente entre las adquisiciones de las personas si se considera que casi un 20% de los clientes cuenta con este artículo, y que en promedio los individuos poseen pocos créditos de este tipo. Esto puede tener su origen en que los créditos hipotecarios suelen estar relacionados a montos elevados de dinero, por lo que no es un producto que las personas contraten periódicamente.

Pasando a las variables relacionadas al envío y apertura promedio de correos, estas evidencian que típicamente se envían cerca de 17 correos al mes por persona. Como mucho, una persona puede recibir casi 28 correos mensualmente, es decir, recibir aproximadamente 1 correo diario. Sin embargo, la lectura frecuente de un cliente por mes es cercana a 2 correos, evidenciando que la mayoría de las personas se caracterizan por leer pocos mails mensualmente. Más aún, la apertura máxima no supera los 24 correos, reforzando la brecha que existe entre el número de correos que se envían y la cantidad de correos que se leen mes a mes por parte de los clientes. Lo señalado previamente se puede apreciar en el siguiente gráfico de cajas:

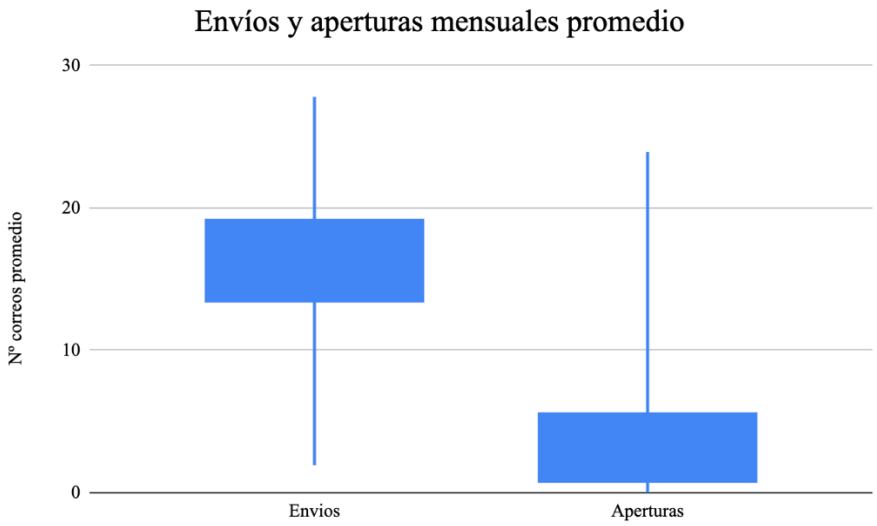

Figura 10: Comparación entre envíos y aperturas mensuales promedio.
Fuente: Elaboración propia con datos de la Base de personas.

En cuanto a los datos sobre la lectura de los correos por parte de las personas, entre ellos destaca que el martes es el día en que los clientes suelen leer más correos, seguido por el miércoles y el jueves. Por su parte, sábado y domingo presentan el caso opuesto, es decir, los días del fin de semana son los menos preferidos para leer mails. Siguiendo esta lógica con las horas del día, la jornada de la tarde es cuando los clientes generalmente tienen mejores niveles de lectura. Más aún, entre las 13:00 hrs. y 16:00 hrs. se encuentran los momentos en que las personas más correos leen. La situación opuesta ocurre durante la madrugada, en específico entre las 00:00 hrs. y 03:00 hrs., que es el intervalo del día menos predilecto para leer correos.

Finalizando con la Base de personas, la variable de demora de lectura presenta algunas observaciones indeterminadas cuando los individuos no registran correos leídos. Sin embargo, para la parte determinada de esta variable, se tiene que existen casos en que los clientes demoran, en promedio, menos de 2 minutos en leer los mails, así como también existen casos en que las personas demoran meses en leer los correos. De esta forma, la variable mencionada tiene una media que se encuentra entre los 3 y 4 días.

La segunda base que se construye, nombrada como “Base de correos”, contiene información sobre los correos enviados a los 405.312 clientes que posee la Base de personas. En este caso, el intervalo temporal utilizado comprende los meses entre octubre del año 2020 y octubre del 2021, nuevamente sin tener en cuenta el mes de febrero del 2021 por la cantidad de datos que presenta en la Base de envíos. Así, la Base de correos posee cerca de 80 millones de filas, donde cada una representa un correo enviado. En la tabla a continuación se pueden ver las variables que contiene esta base según su tipo:

Tabla 6: Variables presentes en Base de correos.
Fuente: Elaboración propia.

Variables tipo *"string"*	Variables tipo *"integer"*
Día de envío	ID
Ámbito del correo	Abierto
Tipo de correo	Hora de envío

En el siguiente listado se describe cada una de las variables presentadas en la **Tabla 6**:

- Día de envío: Abreviación de 3 letras que indica el día en que se envió un correo (en inglés). Por ejemplo, si un correo se envió un día lunes, Día de envío = "Mon".
- Ámbito del correo: Hace referencia al objetivo de un correo enviado. Posee los siguientes niveles:
    - Fidelizar.
    - Vender.
    - Informar.
    - Cobrar.
    - Atraer.
    - Pyme.
- Tipo de correo: Se relaciona con la información contenida en el correo. Posee los siguientes niveles según el ámbito del mail[18]:
    - Fidelizar:
        - Actualizar datos.
        - Habilitación.
        - Inversiones.
        - Onboarding.
        - PAT.
        - Tarjeta.
    - Vender:
        - Aumento cupo.
        - Avance.
        - Consumo.
        - Cuotización.
        - Hipotecario.
        - Inversiones.
        - Seguro.
        - Tarjeta.
    - Informar:
        - Canales digitales.
        - News.
        - Otros.

[18] Aquí se presentan las combinaciones entre ámbito y tipo de correo, es decir, las clases de correos.

    - Cobrar:
        - Riesgo.
    - Atraer:
        - Planes.
    - Pyme:
        - Pyme.
- ID: Número entero que reemplaza el RUT de cada persona. Va desde 1 a 405.312.
- Abierto: Indicador binario que toma el valor 1 si el correo fue abierto, 0 cuando no.
- Hora de envío: Número entero que se relaciona con la hora en que se envió un correo. Por ejemplo, si un correo se envía entre 11:00:00 y 11:59:59 hrs., Hora de envío = 11.

La Base de correos presenta entre sus datos aproximadamente 18,6 millones de correos que registran una apertura. Analizando la apertura de correos según su ámbito, se obtienen los resultados presentados en la siguiente tabla:

Tabla 7: Apertura de correos según ámbito.
Fuente: Elaboración propia con datos de la Base de correos.

Ámbito	Nº correos enviados [Millones]	Porción de correos abiertos
Fidelizar	42,64	23,56%
Vender	13,70	27,83%
Informar	19,24	24,11%
Cobrar	0,17	35,22%
Atraer	0,03	28,20%
Pyme	0,03	28,28%

La tabla anterior muestra que los correos del ámbito "Cobrar" tienen un ratio de apertura mayor que los demás, lo cual puede fundamentarse en que los mails que tratan de deudas suelen generar una preocupación en las personas. De esta forma, cuando los clientes reciben un correo de cobranza, tienen un interés mayor por conocer sobre esta misma. No obstante, cabe mencionar la importancia de los ámbitos "Fidelizar", "Vender" e "Informar" que representan la mayoría de los correos que el banco envía, similar a lo que se evidencia en la sección "Desarrollo metodológico: Comprensión de los datos".

En el caso de los tipos de correo, destaca que "Tarjeta", "Canales digitales" y "Otros" sean los más enviados con 39,40 millones, 10,08 millones y 6,74 millones de registros respectivamente; mientras que los tipos "Riesgo", "Avance" y "Onboarding" sean los que presenten mayor ratio de apertura con 35,18%, 34,64% y 32,07% correspondientemente. La cantidad total de correos enviados y los niveles de lectura para todos los tipos de correo se encuentra en el **Anexo A**.

Ahora, estudiando la lectura de correos respecto al día de envío, se obtienen los siguientes resultados:

Tabla 8: Apertura de correos según día de envío.
Fuente: Elaboración propia con datos de la Base de correos.

Día	Nº correos enviados [Millones]	Porción de correos abiertos
Lunes	14,37	25,19%
Martes	17,97	25,57%
Miércoles	15,72	24,14%
Jueves	13,62	23,66%
Viernes	13,63	23,74%
Sábado	0,23	13,51%
Domingo	0,26	28,12%

Las cifras mostradas en **Tabla 8** reflejan que los días martes, miércoles y domingo podrían considerarse como los mejores días para enviar correos, ya que presentan mayores ratios de apertura de mails, sobre todo el día domingo. Sin embargo, hay que tener en cuenta que, según los datos de esta misma base, los fines de semana de semana suelen realizarse menos envíos. Así mismo, se evidencia un notorio uso de los días de la semana para mandar los mails, lo que se relaciona con las restricciones internas del banco, que establecen no mandar correos los fines de semana.

En cuanto a las horas de envío, se tiene que el intervalo entre las 12:00:00 hrs. y 14:59:59 hrs. presenta las horas en que más correos se envían, sumando 37,81 millones de registros. En otras palabras, entre esas 3 horas acumulan cerca del 50% de los datos registrados. Además, llama la atención que algunas horas presentan ratios de lectura bastante altos: casi 75% entre las 01:00:00 hrs. y 01:59:59 hrs., 61,54% entre las 18:00:00 hrs. y 18:59:59 hrs., y cerca de 56% entre las 21:00:00 hrs. y 21:59:59 hrs. y entre las 22:00:00 hrs. y 22:59:59 hrs. Sin embargo, estos intervalos horarios presentan muy pocos datos (menos del 0,001% de los datos) como para considerarlos válidos en este caso. Por ende, tomando en cuenta resultados con un mayor número de observaciones, resulta que los intervalos 17:00:00 hrs. a 17:59:59 hrs., 04:00:00 hrs. a 04:59:59 hrs., 06:00:00 hrs. a 06:59:59 hrs. y 12:00:00 hrs. a 12:59:59 hrs., presentan tasas de lectura mayores: 26,52%, 25,61%, 25,40% y 25,33% respectivamente. El número de mails enviados y las tasas de lectura para todas las horas se puede ver en el **Anexo B**, donde se evidencia otra parte de estas restricciones internas del banco, que en este caso estipulan no mandar mails fuera de la jornada laboral, o sea, acotar los envíos entre las 07:00 hrs. y 19:00 hrs. aproximadamente.

Para finalizar, es importante mencionar que en las dos bases descritas en esta sección, no se consideran registros de correos de los grupos Rebote y Erróneos, detallados en la sección "Desarrollo metodológico: Comprensión de los datos". Mientras que aquellos correos del grupo Desuscripción, no se contabilizan en el número de correos abiertos pero si en la cantidad de mails enviados.

7.3.3 Estudio preliminar de saturación

En esta parte comienza el estudio del grado de saturación para los individuos que conforman la Base de personas, a través del criterio "Caída en la tasa de apertura". Para realizar aquello, en primer lugar, el intervalo temporal utilizado en la base mencionada es dividido 4 veces en 3 periodos de 4 meses, con el fin de disminuir la estacionalidad que puedan presentar ciertas épocas del intervalo temporal trabajado. Así, por ejemplo, la primera división de los 15 meses en estudio incluye en su primer periodo los meses de julio, agosto, septiembre y octubre del año 2020; el segundo periodo contiene los meses de noviembre y diciembre del 2020, y enero y marzo del 2021; mientras que el tercer periodo está compuesto por los meses de abril, mayo, junio y julio del año 2021. A medida que se avanza en las divisiones, los meses de cada periodo se cambian por el mes siguiente respectivo. Para revisar el detalle de los meses que considera cada periodo en cada división, véase el **Anexo C.**

Posteriormente, a cada persona se le calcula una tasa de apertura mensual promedio en cada uno de los periodos que comprenden las divisiones. Estas tasas promedio se calculan mediante el promedio simple entre las tasas mensuales de apertura de los meses que componen cada periodo. Lo anterior se refleja en la expresión a continuación:

$$T_{p,i,j} = \frac{\sum_k T_{p,i,j,k}}{4} \quad (17)$$

Donde *p* indica la persona a la cual se le calcula su tasa promedio, *i* corresponde al periodo, *j* señala la división, *k* son los meses que comprende cada periodo y *T* la tasa de apertura mensual.

Una vez que se obtienen estas tasas de apertura promedio para todas las personas en los 12 periodos (4 divisiones con 3 periodos), se calcula una media de estas tasas resultantes, pero esta vez, según la división. Nuevamente se utiliza el promedio simple, con lo que la fórmula de este caso se ve de la siguiente forma:

$$T_{p,i} = \frac{\sum_j T_{p,i,j}}{4} \quad (18)$$

Así, cada persona posee una tasa mensual de apertura relacionada a un bloque o conjunto de 4 periodos: el primer bloque, formado por los primeros periodos, se considera como "Histórico", donde la tasa de apertura resultante se interpreta como el nivel de lectura natural de las personas; el segundo bloque, conformado por los segundos periodos, se designa para estudiar la "Caída" que tengan las tasas de apertura al comparar con el bloque anterior; finalmente el tercer bloque, compuesto por los terceros periodos, se deja para identificar "Recuperaciones" en las tasas de apertura frente a las caídas que pueden darse en el bloque previo.

Con lo anterior, se clasifican a las personas en tres grupos según la variación que tengan sus tasas de apertura promedio a través de estos 3 bloques: "Sube", donde están consideradas las personas que aumentan en más de 5 p.p. su tasa promedio de apertura entre un bloque y el siguiente; "Mantiene", que toma en cuenta aquellos individuos que no cambian en más de 5 p.p. sus tasas promedio entre dos bloques contiguos; y "Baja", donde se agrupan los clientes que disminuyen en más de 5 p.p. sus tasas promedios al comparar un bloque con el anterior. Los resultados de esta clasificación se muestran en el siguiente esquema[19]:

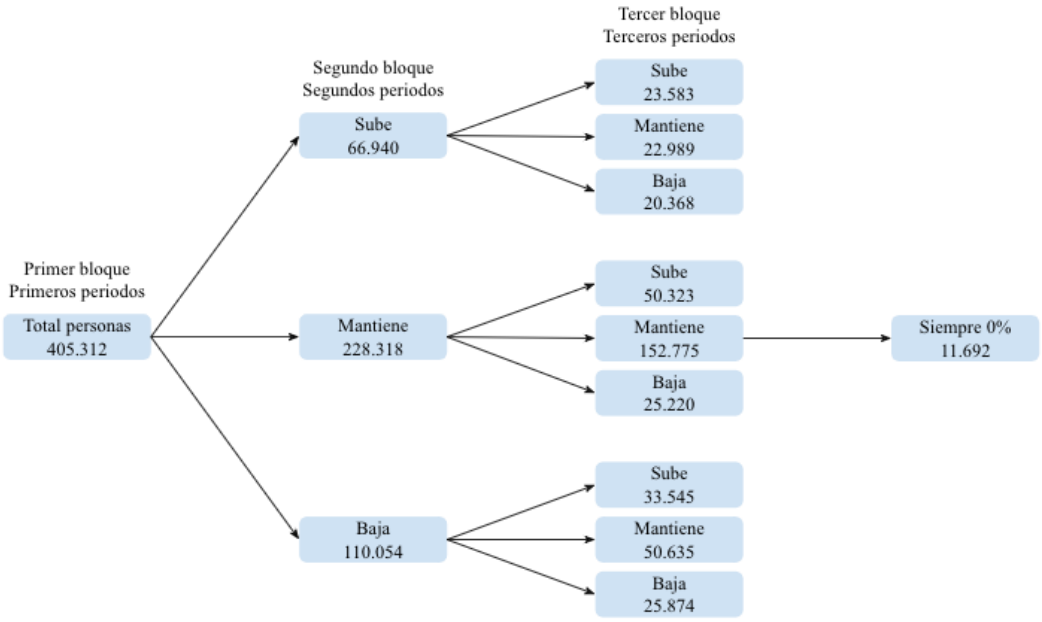

Figura 11: Cambio en tasas mensuales de apertura promedio.
Fuente: Elaboración propia.

La figura anterior muestra que de las 405.312 personas que forman parte de la Base de personas, existen 110.054 que bajaron su nivel de lectura promedio de correos entre el primer y el segundo bloque de periodos. Recordando el criterio de saturación escogido, este grupo de más de 110 mil clientes, cumplen con la característica principal para ser considerados como saturados (tener una disminución en el *open rate*), sobre todo las más de 70 mil personas que para el tercer bloque no logran aumentar o recuperar su tasa de lectura. Además, destaca el hecho de que existen cerca de 12 mil personas que entre julio del 2020 y octubre del 2021 tuvieron 0% de lectura, las cuales fácilmente podrían considerarse como saturados también.

Con el análisis explicado previamente, se elaboran los siguientes gráficos para ejemplificar la evolución en las tasas de apertura de algunas personas que podrían reconocerse como saturadas y no saturadas:

[19] Al ser un estudio preliminar, la cota de 5 puntos porcentuales usada en esta parte se escoge de manera arbitraria.

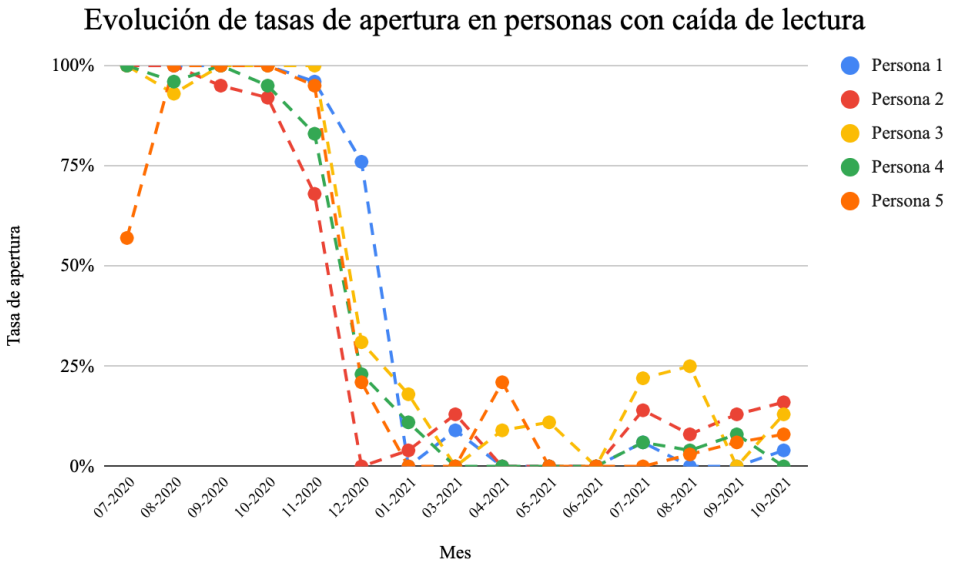

Figura 12: Caídas en tasas mensuales de apertura.
Fuente: Elaboración propia con datos de la Base de envíos.

El gráfico anterior muestra los cambios en el tiempo de las tasas de apertura mensual para aquellas personas que evidencian una caída de estas en el segundo bloque de periodos. Además, para el tercer bloque, este grupo refleja una nula recuperación, por lo que podría concluirse que estos individuos están saturados de los correos electrónicos que envía el banco.

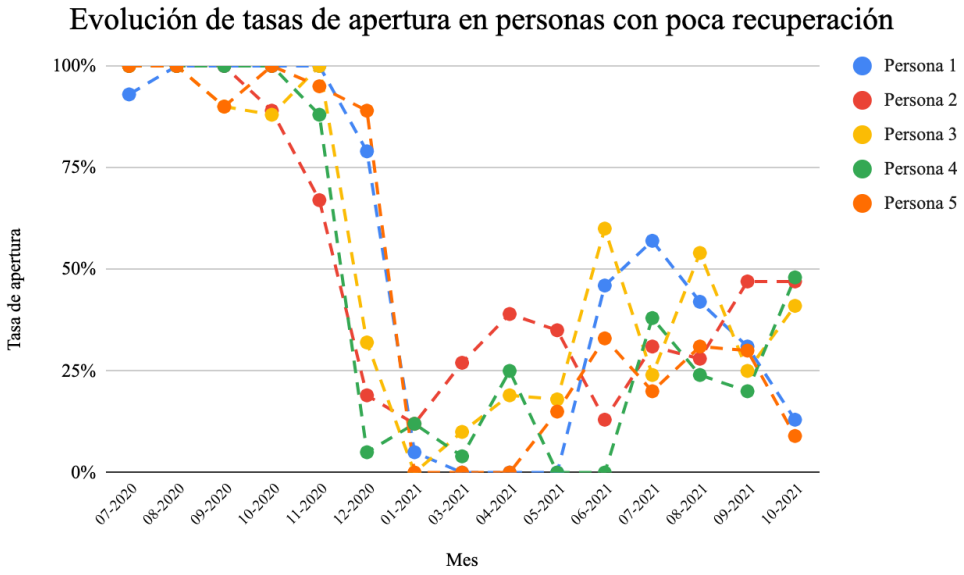

Figura 13: Poca recuperación en tasas mensuales de apertura.
Fuente: Elaboración propia con datos de la Base de envíos.

Similar a lo mostrado en la **Figura 12**, la **Figura 13** refleja la evolución de las tasas de lectura para personas que tienen una caída en el segundo bloque de periodos, y que para el tercer bloque no

logran recuperarse lo suficiente, por lo que nuevamente podría decirse que estas personas están saturadas por el email marketing de la empresa.

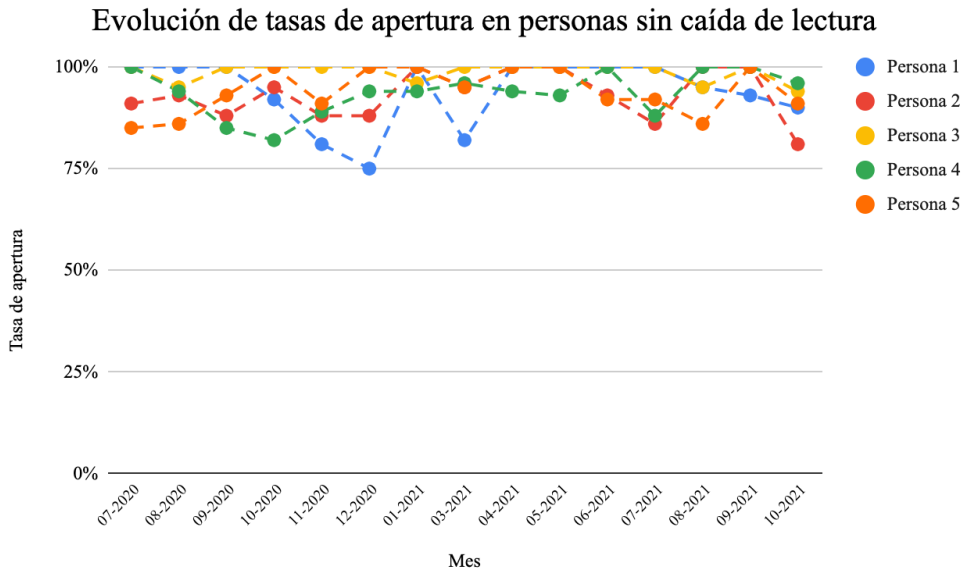

Figura 14: Altas tasas mensuales de apertura.
Fuente: Elaboración propia con datos de la Base de envíos.

Por otro lado, la **Figura 14** muestra tasas de apertura para personas que se caracterizan por leer la mayoría de los correos que reciben por parte del banco. Más aún, destaca el hecho de que este grupo no evidencia caídas en sus niveles de lectura, por lo que este tipo de individuos podría considerarse como no saturados.

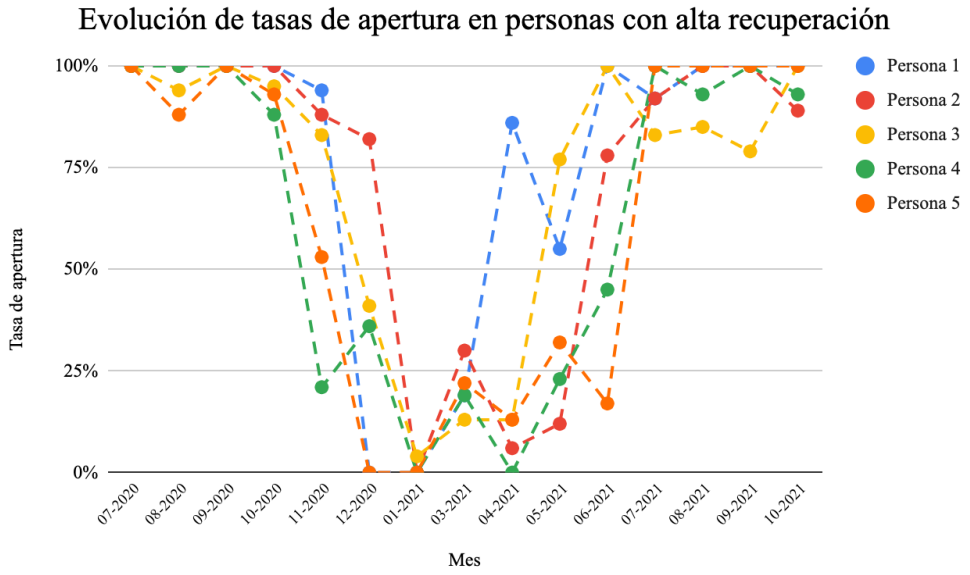

Figura 15: Alta recuperación en tasas mensuales de apertura.
Fuente: Elaboración propia con datos de la Base de envíos.

Distinto a lo que se puede ver en la **Figura 13**, la **Figura 15** refleja tasas de apertura en personas que tienen un buen nivel de recuperación, frente a una caída evidenciada en el segundo bloque de periodos. Más aún, las personas de este grupo alcanzan en el tercer bloque tasas de lectura similares a las que tenían un principio, por lo que no es errado pensar que este tipo de personas no están saturadas por el email marketing del banco.

7.3.4 Exigencias de caídas y recuperaciones

Luego del análisis anterior, es necesario establecer exigencias en las caídas y las recuperaciones de las tasas de apertura para definir la saturación. En otras palabras, se debe establecer desde qué punto una disminución en las tasas de lectura se considera como caída, y desde qué punto un aumento de las mismas, luego de evidenciar una caída, se toma como recuperación.

Para esto, nuevamente se consideran las tasas de apertura calculadas por bloques de periodos según lo explicado en la parte previa. Así, las exigencias de las caídas se estudian a través de la "Matriz inicio/caída" considerando a las 110.054 personas que evidencian una disminución en su nivel de lectura entre el primer y segundo bloque. La matriz mencionada se puede ver en **Anexo D**.

Esta matriz muestra en el eje vertical los intervalos de tasas de lecturas que las personas promedian en el primer bloque de periodos, mientras que en el eje horizontal están los intervalos de los niveles de caídas que se evidencian para el segundo bloque. Por ejemplo, si un individuo promedia en el primer bloque un 85% de apertura mensual y en el segundo un 50%, significa que la tasa de lectura disminuye en 35 p.p. y por ende, para el segundo bloque cae al 58,82% de su nivel de apertura inicial[20]. De esta forma, esa persona estaría en el intervalo inicial [80%,90%[y en el intervalo de caída [50%,60%[, por lo que sería parte de los 266 clientes que se contabilizan en la intersección de estos eventos.

Con esta matriz y a través de un procedimiento similar al utilizado en el método del codo, se procede a buscar los intervalos de caída que pueden ser puntos de inflexión en las curvas de los intervalos de lectura inicial. Esto sirve para tener una idea de en qué nivel de caída se podría establecer la exigencia de estas mismas, para todos los niveles iniciales de apertura. Cabe aclarar que las curvas de esta parte se realizan mediante fracciones acumuladas de personas, las cuales se obtienen al sumar los cocientes entre la cantidad de individuos que están en cierto intervalo inicial de lectura y en un determinado nivel de caída, y la cantidad total de clientes que están en ese intervalo inicial de apertura. Así, los resultados se presentan a continuación:

[20] 50 es el 58,82% de 85.

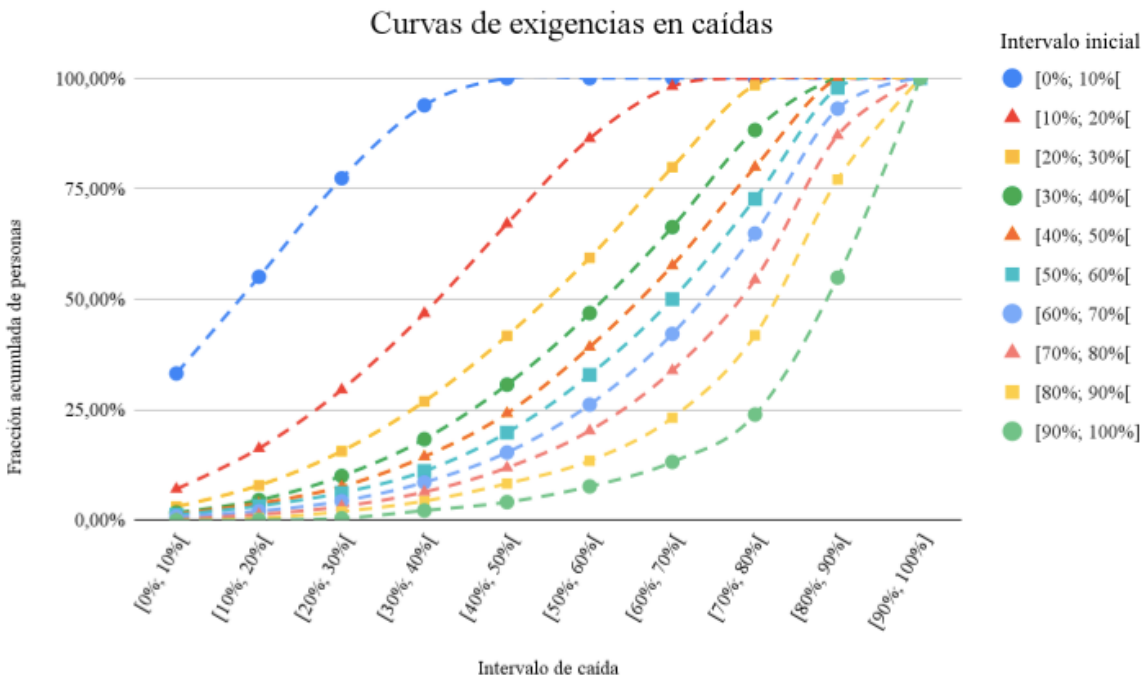

Figura 16: Inflexiones para exigencias de caídas.
Fuente: Elaboración propia.

En la **Figura 16** se puede apreciar que para la mayoría de los intervalos de lectura inicial, los puntos de inflexión están bajo el 50% de fracción acumulada de personas. Si bien la figura no muestra un punto de inflexión claro en la mayoría de las curvas, este método da una primera aproximación del nivel que podría considerarse para establecer la exigencia de las caídas. Un caso de esto es la curva del intervalo inicial [90%,100%], la cual pareciera tener su inflexión entre los intervalos de caída [60%,70%[y [70%,80%[, por ende, la búsqueda de la exigencia de caída para ese nivel inicial de lectura, se acota a estudiar en profundidad entre esos niveles de caídas.

Para las exigencias en las recuperaciones, se realiza un procedimiento análogo al que se detalló anteriormente para las caídas. Sin embargo, en este caso se considera el grupo de personas que tiene un grado de recuperación en sus tasas de lectura para el tercer bloque de periodos, frente a una caída evidenciada en el segundo. Es decir, se consideran los 33.545 individuos que suben su nivel de lectura para el tercer bloque luego de que en el segundo tuvieron una disminución de este mismo nivel. Así, se elabora la "Matriz inicio/recuperación" que se puede revisar en el **Anexo D**.

Similar a la Matriz inicio/caída, la Matriz inicio/recuperación tiene en su eje vertical los niveles de lectura que las personas promedian en el primer bloque de periodos, y en su eje horizontal están los niveles de recuperación que los mismos clientes tienen para el tercer bloque, siempre y cuando hayan evidenciado una caída en el segundo. Por ejemplo, si un individuo en el primer bloque promedia un 70% de apertura mensual, en el segundo 35% y en el tercero 60%, significa que la tasa de lectura se recupera en un 85,71% respecto al nivel inicial[21]. Así, esta persona estaría

[21] 60 es el 85,71% de 70.

asignada al intervalo de apertura inicial [70%,80%[y en el intervalo de recuperación [80%,90%[, por lo que sería uno de las 279 clientes que se registran en la intersección de estos intervalos.

Nuevamente, aplicando un proceso análogo al método del codo, se buscan los intervalos de recuperación que podrían ser puntos de inflexión para todas las curvas de los intervalos de apertura inicial. Realizando un procedimiento similar al explicado en las exigencias de caídas, se obtienen los siguientes resultados:

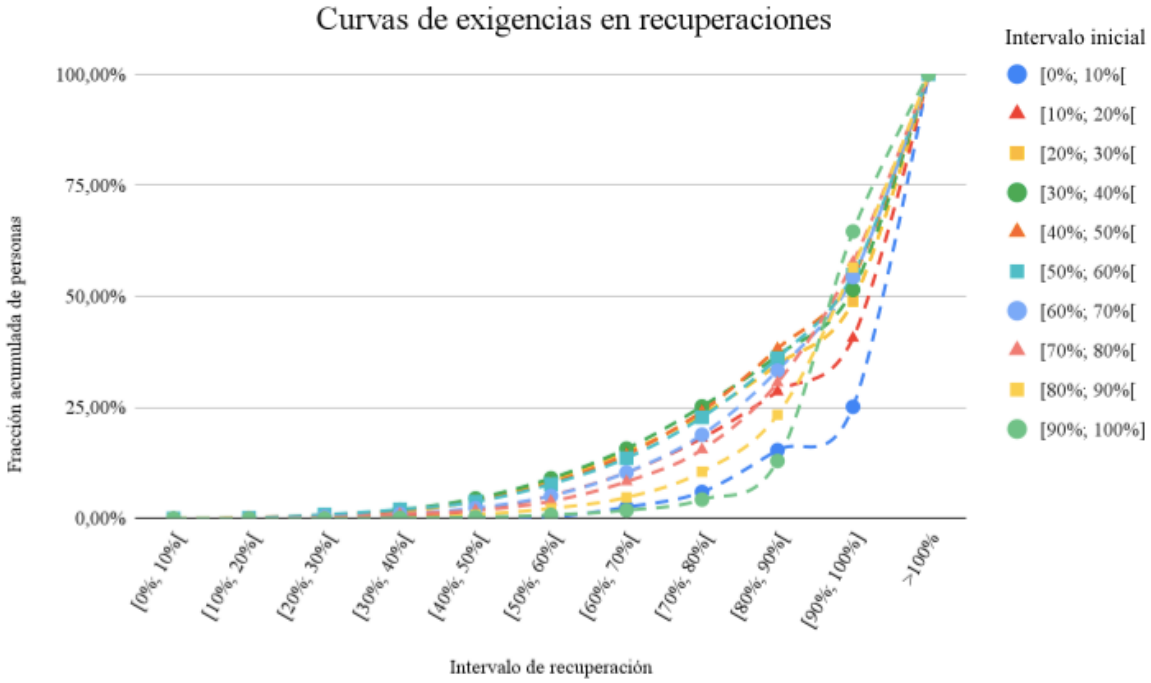

Figura 17: Inflexiones para exigencias de recuperaciones.
Fuente: Elaboración propia.

Parecido a lo obtenido para las caídas, los puntos de inflexión en este caso, se ubican, en su mayoría, debajo del 50% de fracción acumulada de personas. Sin embargo, estas inflexiones se encuentran en intervalos de recuperación mayores que los que se reflejan en el caso de las caídas, lo cual hace sentido considerando que las recuperaciones buscan que las tasas de lectura aumenten para el tercer bloque de periodos teniendo en cuenta las caídas que sufren en el segundo.

Análogamente, estos procedimientos para encontrar exigencias en los niveles de caídas y recuperaciones se realizaron con datos de clientes que se han desuscrito de las campañas de email marketing del banco, pero la cantidad de personas que cumplen con las restricciones usadas no son suficientes para realizar el análisis (~600 registros), obteniéndose resultados más distorsionados y menos concluyentes.

A partir de los análisis realizados en esta parte, se buscan en las curvas presentadas en las **Figura 16** y **Figura 17**, aquellos puntos de inflexión que puedan servir como exigencias para caídas y

recuperaciones. Con esto, se establecen las cifras que indican si ciertas disminuciones o aumentos en las tasas de lectura se consideran en efecto como caídas o recuperaciones respectivamente. Los resultados se presentan en la tabla a continuación:

Tabla 9: Exigencias para caídas y recuperaciones según intervalo de lectura inicial.
Fuente: Elaboración propia.

Intervalo inicial	Caída	Recuperación
[90%,100%]	73%	81%
[80%,90%[	66%	81%
[70%,80%[	56%	85%
[60%,70%[	48%	83%
[50%,60%[	38%	73%
[40%,50%[	39%	69%
[30%,40%[	31%	75%
[20%,30%[	21%	72%
[10%,20%[	10%	73%
[0%,10%[	1%	79%

La tabla anterior muestra en su columna izquierda los diferentes intervalos iniciales de lectura que se presentan en el primer bloque de periodos en estudio. En la columna central se encuentran las exigencias, para diferentes niveles iniciales de apertura, que determinan si disminuciones de estos niveles iniciales hacia el segundo bloque de periodos son consideradas como caídas. Análogamente, en la columna derecha están las cifras que definen aquellas alzas en las tasas de lectura, que una vez evidenciada una caída, se denominan como recuperaciones en el tercer bloque de periodos.

Cabe aclarar que en el caso de las caídas, la **Tabla 9** presenta un límite superior para las disminuciones en los niveles de lectura. Por ejemplo, si una persona promedia inicialmente 95% de apertura de correos (se encontraría en el intervalo inicial [90%,100%]) y para el segundo bloque de periodos su lectura baja a 69,35%, esta disminuyó al 72,99% respecto a su nivel de apertura inicial, por lo que este descenso se consideraría como caída. En cambio, si para el segundo bloque su lectura bajase a 69,35%, esta disminuiría al 73% en relación a su nivel inicial de apertura, por lo que no sería una caída.

En el caso de las recuperaciones, la tabla muestra un límite inferior para las alzas de lectura que se pueden dar en el tercer bloque de periodos luego de evidenciar una caída en el bloque previo. Por ejemplo, si una persona promedia inicialmente 66% de lectura (se encontraría en el intervalo inicial [60%,70%[), en el segundo bloque su apertura cae a 40%, y para el tercer bloque de periodos promedia 54,78% de lectura, su apertura crecería en este último periodo hasta el 83% respecto a su nivel inicial, por lo que este caso se trataría de una recuperación. Por el contrario, si para el tercer bloque de periodos su apertura promedio es de 54,77%, esta aumentaría para este bloque hasta el 82,98% en relación al nivel inicial, por lo que no se podría considerar como una recuperación.

Los resultados mostrados en la **Tabla 9** evidencian que las exigencias de caídas disminuyen al mismo tiempo que bajan los niveles iniciales de lectura. De esta forma, mientras menor sea el nivel de apertura inicial, la disminución de esta para el segundo bloque de periodos debe ser cada vez más abrupta para que se considere una caída. Más aún, destaca el hecho de que para aquellas personas que promedien inicialmente un nivel de lectura del intervalo [0%,10%[, la caída se evidencia si en el segundo bloque prácticamente dejan de leer correos. Situación contraria ocurre con las exigencias de las recuperaciones, las cuales se mantienen cerca del 70% u 80% para todo nivel inicial de lectura.

Finalmente, según las exigencias para las caídas y las recuperaciones detalladas, existen 23.377 personas que son consideradas como saturadas, lo cual se agrega como una variable binaria denominada "Saturado" en la Base de personas, de forma tal que el valor 1 indica si una persona esta saturada, 0 si no.

7.4 Modelación y evaluación

En esta sección del trabajo se explican los procedimientos realizados para llevar a cabo los modelos matemáticos presentados en las secciones "Marco conceptual: Segmentación" y "Marco conceptual: Predicción".

La modelación se realiza a través del programa computacional *RStudio*, el cual no cuenta con la capacidad para incorporar las bases de datos por completo en los modelos. Por ende, en esta parte también se explica la obtención de muestras de datos mediante la técnica de Muestreo aleatorio simple.

A medida que se obtienen resultados, es posible comparar los modelos utilizados. Así, la evaluación también se incluye en esta parte, donde se hace uso de la mayoría de las métricas propuestas en la sección "Marco conceptual: Criterios de evaluación".

7.4.1 Segmentación

En esta parte se explica el proceso con el cual se segmenta el conjunto de clientes estudiados en este trabajo, para así diseñar las políticas de toques según estos grupos. En un principio, el *clustering* se realiza para un número menor de personas mediante las técnicas de Agrupamiento jerárquico aglomerativo y *K*-medias. Posteriormente, los resultados obtenidos se utilizan para entrenar el modelo de *K*-vecinos más cercanos, y así clasificar a cada uno de los clientes en el segmento que le corresponda.

De esta manera, antes de comenzar con la modelación de los segmentos de clientes, es necesario realizar un muestreo de la Base de personas. Dada la capacidad de *RStudio* para realizar un *clustering*, las muestras obtenidas en esta parte de la modelación poseen poco más de 10 mil

observaciones (~2,5% del total de datos). Según esto, se extraen 10 muestras no excluyentes[22] para revisar el flujo de la segmentación en diferentes grupos de datos de la misma población. Para obtener cada una de las muestras, se cuidan las distribuciones de las variables de género, estado civil, nivel educacional y saturación. A continuación, se presenta el reparto de niveles de estas variables categóricas en la Base de personas para así hacer una idea de la composición de las muestras:

Tabla 10: Distribución de variables categóricas en Base de personas.
Fuente: Elaboración propia.

Género		**Saturación**	
M	F	Saturado	No saturado
59,20%	40,80%	5,77%	94,23%

Estado civil					
CAS	CCV	DIV	SEP	VIU	SOL
37,41%	0,31%	0,14%	6,71%	1,09%	54,34%
Nivel educacional					
BAS	EUN	MED	SINE	TEC	UNV
0,21%	11,16%	14,09%	0,01%	30,44%	44,08%

Dados los resultados de la **Tabla 10**, se procura que las variables categóricas expuestas anteriormente, para las 10 muestras resultantes, no varíen en más de 5 p.p. en todos los niveles.

Posterior a la obtención de muestras, es necesario definir las variables a utilizar en los modelos de segmentación. En un principio, se incorporan en el *clustering* todas las variables que posee la Base de personas, no sin antes ser normalizadas[23] según los datos de cada muestra, a través de la siguiente fórmula[7]:

$$z_{i,k} = \frac{x_{i,k} - min(x_k)}{max(x_k) - min(x_k)} \quad (19)$$

Donde:

- $z_{i,k}$: Dato normalizado para la observación *i* en la variable *k*.
- $x_{i,k}$: Dato real de la observación *i* en la variable *k*.
- x_k: Datos de la variable *k*.

De esta forma, mediante el Método de la silueta para Agrupamiento jerárquico aglomerativo y *K*-medias, se verifica que el óptimo de *clusters* a considerar en ambos modelos está cerca de los 20 segmentos para la mayoría de las muestras, provocando que las segmentaciones generen varios grupos de pocas observaciones (con menos de un 1% de los datos de la muestra respectiva).

[22] Es decir, una persona puede aparecer en más de una muestra.
[23] Los datos quedan expresados en el rango [0,1].

Dado lo anterior, se revisa la composición de los segmentos resultantes, con lo que se puede concluir que, al incorporar varias variables categóricas en los modelos de *clustering*, estos procuran diferenciar a los individuos principalmente por los niveles que presenten en este tipo de variables. El caso más típico era obtener un par de segmentos con características muy similares, donde la principal diferencia era que uno estaba compuesto sólo por hombres, mientras que el otro agrupaba sólo mujeres. Es en este punto donde se decide dejar de lado la mayoría de las variables categóricas que presentan los datos, ya que el foco de esta parte del trabajo está sobre los niveles de lectura de correos. Por ende, un individuo es considerado buen lector independiente de su género, estado civil o nivel educacional, entre otros.

En cuanto a las variables numéricas, se descartan los datos de edad y antigüedad, dado que las segmentaciones comentadas anteriormente mostraban que los *clusters* tenían un promedio de edad y antigüedad que se acercaba bastante a la media poblacional, por lo que se intuye que estas variables no caracterizan los segmentos resultantes. Además, las variables Promedio mensual de envíos y Promedio mensual de aperturas no se consideran ya que se incorporan en los modelos a través de la tasa de lectura promedio.

De esta forma, entre las variables consideradas en el *clustering* se encuentran aquellas que indican la cantidad de cada producto que tienen los individuos, las cuales se interpretan como un grado de fidelidad de las personas con el banco; y las que se obtienen gracias a los niveles de lectura de las personas, tales como las variables Saturado, Demora promedio de lectura y Promedio mensual de *open rate.*

Escogidas y normalizadas las variables a utilizar en los modelos de segmentación, es necesario definir el número óptimo de *clusters* a considerar. Para esto, se utiliza el Método de la silueta en un rango de [1,10] segmentos, y así se evalúan los coeficientes de silueta promedio para las 10 muestras en los modelos de Agrupamiento jerárquico aglomerativo y *K*-medias. En el **Anexo E** se pueden revisar los coeficientes de silueta obtenidos para cada muestra en los dos modelos de *clustering*.

Según los resultados mostrados en el **Anexo E**, se calcula una silueta promedio para ambos modelos de segmentación. Esto se realiza mediante el promedio simple de los coeficientes de silueta obtenidos en cada una de las muestras, según el número de *clusters* correspondiente. Este resultado se muestra a continuación:

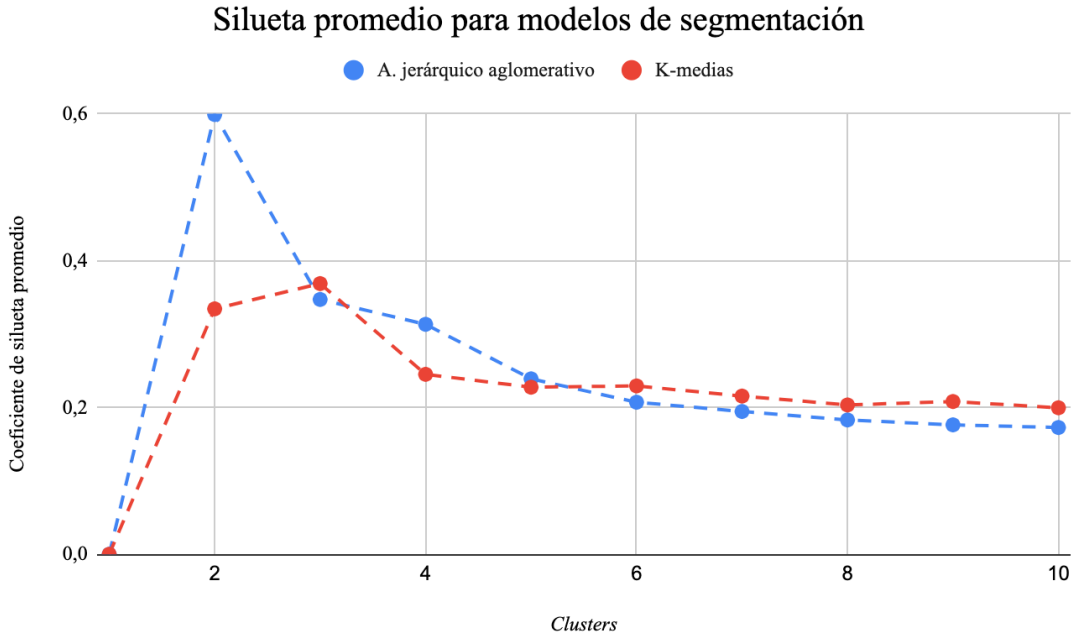

Figura 18: Curva de silueta promedio para Agrupamiento jerárquico aglomerativo y *K*-medias.
Fuente: Elaboración propia.

La figura anterior muestra que para una segmentación realizada con Agrupamiento jerárquico aglomerativo, la cantidad óptima de *clusters* a considerar es 2. Análogamente, para el caso de *K*-medias, el número óptimo de segmentos a utilizar es 3. Si bien esto presenta una cantidad óptima de *clusters* bajo el fundamento del coeficiente de silueta, se hace interesante revisar el flujo de los modelos de segmentación para conocer cómo se van formando los grupos.

Cabe destacar también que la **Figura 18** muestra que, en el rango de [1,5] segmentos, el método de Agrupamiento jerárquico aglomerativo generalmente tiene un mejor desempeño al compararlo con *K*-medias. Sin embargo, ocurre lo contrario en el intervalo de [6,10] *clusters*, aunque en esta segunda mitad del gráfico se evidencia una mayor paridad entre el ajuste que presentan ambos modelos. Esto podría dar una idea del modelo que conviene utilizar finalmente, según la cantidad de segmentos que se busque con el *clustering*.

Según lo señalado anteriormente, se llevan a cabo las segmentaciones en un rango de [2,8] *clusters*, para las 10 muestras trabajadas, mediante ambos métodos mencionados. Esto se realiza para revisar cómo fluye el *clustering*, y por ende, comprobar cómo se van conformando los segmentos. Cabe mencionar que el intervalo de segmentos utilizado en este caso, se establece teniendo en cuenta que la **Figura 18** pareciera mostrar que más allá de los 8 segmentos, no hay mayor mejora en el coeficiente de silueta promedio.

Así, el método de Agrupamiento jerárquico aglomerativo va formando los grupos tal como muestra el esquema a continuación:

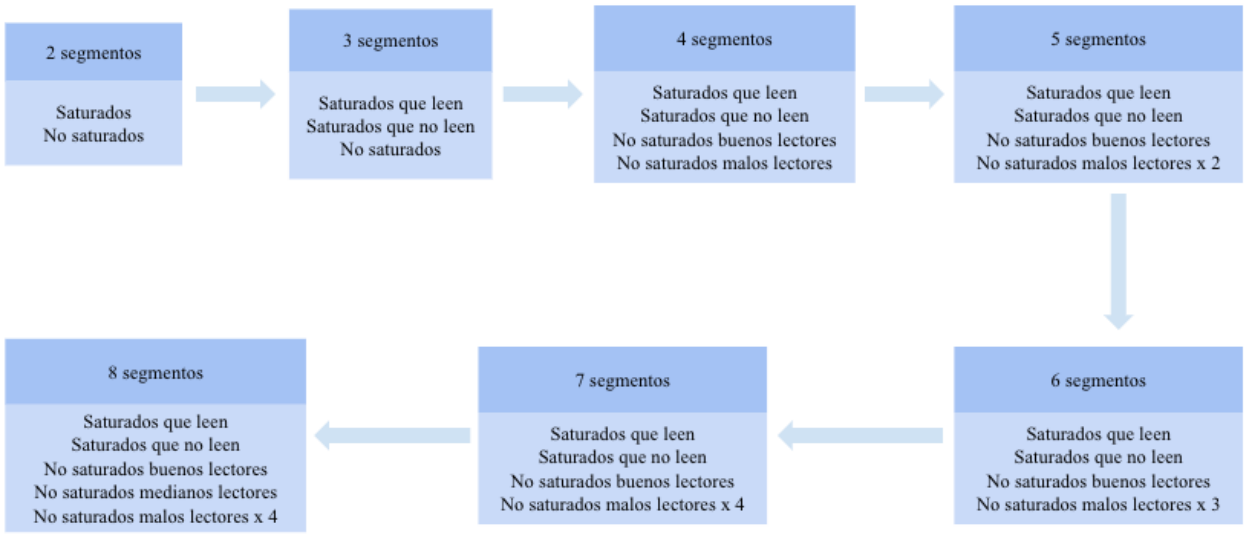

Figura 19: Flujo de segmentación a través de Agrupamiento jerárquico aglomerativo.
Fuente: Elaboración propia.

A continuación se describen los grupos mostrados en la **Figura 19**:

- Saturados: Personas que su variable de saturación es 1.
- No saturados: Individuos que su variable de saturación es 0.
- Saturados que leen: Segmento de personas que se consideran saturadas y que promedia una tasa de lectura mayor a 0%.
- Saturados que no leen: Grupo de personas saturadas que tiene un *open rate* promedio de casi 0%.
- No saturados buenos lectores: Conjunto de individuos no saturados que tiene una tasa de apertura media sobre el 60%.
- No saturados medianos lectores: Grupo de personas que no se consideran saturadas y que tiene un *open rate* promedio entre 30% y 60%
- No saturados malos lectores: Segmento de personas no saturadas que posee, en promedio, una tasa de lectura menor al 20%.

Para el caso de la segmentación por *K*-medias, el flujo del *clustering* queda de la siguiente forma:

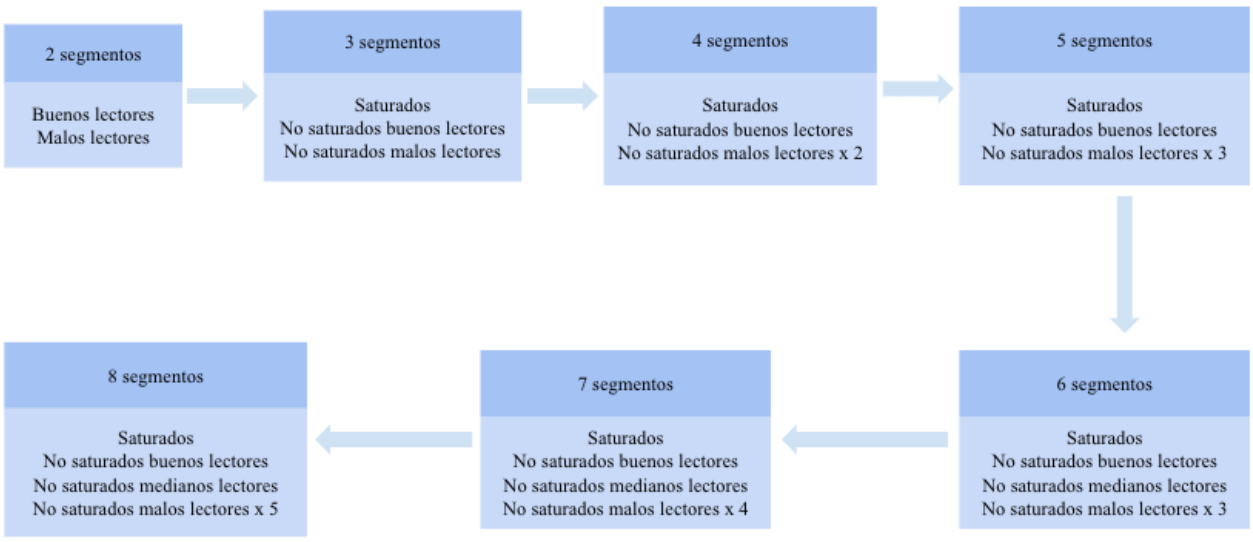

Figura 20: Flujo de segmentación a través de *K*-medias.
Fuente: Elaboración propia.

Con[24]:

- Buenos lectores: Grupo de personas que, en promedio, tiene un *open rate* sobre el 60%.
- Malos lectores: Conjunto de individuos que promedia menos de un 20% en su tasa de lectura.

Con los esquemas presentados en las **Figuras 19** y **20** es posible concluir que, en general, ambos modelos entregan segmentaciones similares según las características de los *clusters* obtenidos. Sin embargo, se decide trabajar con el Agrupamiento jerárquico aglomerativo, ya que da la posibilidad de contar con 2 segmentos de personas saturadas, contrario a los resultados entregados por el modelo de *K*-medias, donde todos los individuos saturados forman parte de un mismo grupo, independientemente de sus niveles de lectura.

No obstante, es importante resaltar que al utilizar el método de Agrupamiento jerárquico aglomerativo, se hace necesario realizar un *clustering* con 8 segmentos para obtener un grupo del tipo "No saturados medianos lectores". Dicho esto, la segmentación se realiza con los datos presentes en la séptima muestra, la cual mediante el método seleccionado, permite establecer claramente niveles altos, medianos y bajos de lectura, gracias a la heterogeneidad de las tasas de apertura promedio evidenciada por cada *cluster.*

Así, al segmentar esta muestra a través de Agrupamiento jerárquico aglomerativo, los grupos presentan las siguientes características para la cantidad de personas, saturados y *open rate* promedio por segmento:

[24] El resto de los grupos son análogos a los explicados para la **Figura 19**.

Tabla 11: Segmentación por Agrupamiento jerárquico aglomerativo.
Fuente: Elaboración propia.

Segmento	Nº personas	Porción de personas	Saturados	Porción de saturados	Promedio *open rate* mensual
1	7.172	70,78%	0	0,00%	13,64%
2	285	2,81%	282	98,95%	0,02%
3	895	8,83%	0	0,00%	85,99%
4	267	2,63%	267	100,00%	23,45%
5	1.163	11,48%	0	0,00%	48,45%
6	337	3,33%	0	0,00%	11,88%
7	8	0,08%	0	0,00%	8,24%
8	6	0,06%	6	100,00%	0,00%

La **Tabla 11** evidencia que existen segmentos que agrupan un número reducido de personas, tal como lo que ocurre con los segmentos 7 y 8. Sin embargo, estos *clusters* presentan características parecidas en porcentaje de saturados y tasa de lectura promedio, a los grupos 1 y 2, respectivamente. Por ende, se decide que aquellos individuos que forman parte de los *clusters* 7 y 8, sean asignados a los segmentos 1 y 2 correspondientemente. De forma análoga, el segmento 6 posee una porción de saturados y *open rate* promedio similar al grupo 1, por lo que también es incluido en este *cluster*.

Incorporando estas modificaciones, la segmentación queda de la siguiente forma:

Tabla 12: Segmentación por Agrupamiento jerárquico aglomerativo modificada.
Fuente: Elaboración propia.

Segmento	Nº personas	Porción de personas	Saturados	Porción de saturados	Promedio *open rate* mensual
1	7.517	74,18%	0	0,00%	13,56%
2	291	2,87%	288	98,97%	0,02%
3	895	8,83%	0	0,00%	85,99%
4	267	2,63%	267	100,00%	23,45%
5	1.163	11,48%	0	0,00%	48,45%

Finalmente, esta segmentación realizada para la séptima muestra, se utiliza para entrenar el modelo de *K*-vecinos más cercanos, con lo que se clasifican en los segmentos mostrados en la **Tabla 12,** a todos los individuos que conforman la Base de personas. En la sección "Despliegue: Segmentación" detallada más adelante, se muestran los resultados obtenidos para el *clustering* de los 405.312 individuos en estudio.

7.4.2 Cantidad de correos a enviar

En esta sección del informe se explica el procedimiento con el que se busca establecer una cantidad de toques a realizar mediante las campañas de email marketing del banco, según la persona a la que se esté contactando. Para esto, se hace uso de los modelos Regresión lineal múltiple y Árbol de regresión, y así pronosticar la cantidad de correos que son capaces de leer las personas mensualmente.

Los datos de la Base de personas nuevamente son incorporados en los modelos matemáticos utilizados en esta parte. Más aún, no es necesario realizar ningún muestreo dado que la capacidad de *RStudio* permite usar todos los registros de la base con las técnicas mencionadas.

Dicho lo anterior, se procede a definir las variables a incluir en los modelos. Se decide utilizar el Promedio mensual de aperturas como variable dependiente, ya que se entiende que la cantidad de mails a mandar por persona, serán tantos como la cantidad de correos leídos que se pronostiquen. Por otra parte, entre las variables exógenas se encuentran la mayoría de la información restante que presenta la Base de personas: datos propios de los individuos como la edad, género, estado civil, nivel educacional y antigüedad; los índices que indican los productos que posee cada persona, los cuales se incluyen en este caso teniendo en cuenta que las variables del número de productos por individuo se consideran a través de la segmentación; y el promedio mensual de envíos, que se incorpora en el entendido de que si una persona recibe más correos, tiene la opción de registrar una cantidad mayor de mails abiertos.

Posteriormente, se entrenan los modelos de Regresión lineal múltiple y Árbol de regresión con un 70% de los datos de la Base de personas. De esta forma, los datos de prueba para obtener las predicciones del número de correos que se abren, corresponden al 30% restante de la base. Cabe mencionar que los datos son particionados según la variable de género.

Así, los rangos predichos para la cantidad de lecturas con ambos modelos, a nivel general[25], se presentan a continuación:

Tabla 13: Predicción del número de aperturas y criterios de evaluación.
Fuente: Elaboración propia.

Modelo	Mínimo	Máximo	R^2	MAE	MAPE	RMSE
Reg. lineal múltiple	-1,78	8,34	0,11	3,22	Inf.	4,26
Árbol de regresión	2,64	5,11	0,07	3,29	Inf.	4,34

Desde la **Tabla 13** es posible concluir que el Árbol de regresión presenta un rango más acotado para las predicciones del número de correos que abren las personas. Además, llama la atención que el límite inferior para el intervalo obtenido con la Regresión lineal múltiple, es negativo. Ante esto

[25] Hace referencia a que no se especifica un segmento de personas.

último, en este trabajo, los valores negativos resultantes de las predicciones de aperturas se interpretan como que las personas no abren correos, es decir, se consideran iguales a 0.

La tabla además muestra que el modelo de Regresión lineal múltiple tendría un mejor ajuste a los datos si se compara con el Árbol de regresión, dado que presenta un R^2 levemente mayor. Este hallazgo se replica con el resto de parámetros de error. No obstante, es válido concluir que ambos modelos presentan bajo ajuste y desempeño en sus predicciones, lo que deja entrever una gran heterogeneidad entre los registros de individuos de la Base de personas que afecta en las mediciones predichas.

Ante lo evidenciado en la **Tabla 13**, las predicciones del número de aperturas se realiza por cada segmento de personas siguiendo un procedimiento análogo al explicado anteriormente, con lo que se obtienen los siguientes resultados para el modelo de Regresión lineal múltiple:

Tabla 14: Predicción de aperturas y criterios de evaluación por segmento con Regresión lineal múltiple.
Fuente: Elaboración propia.

Segmento	Mínimo	Máximo	R^2	MAE	MAPE	RMSE
Segmento 1	0,63	3,59	0,04	1,51	2,52	1,93
Segmento 2	0,00	0,01	0,00	0,00	Inf.	0,02
Segmento 3	2,93	21,23	0,74	1,42	0,10	1,72
Segmento 4	-1,43	7,67	0,20	2,40	Inf.	3,03
Segmento 5	7,82	10,34	0,01	2,13	0,26	2,60

Mientras que los datos resultantes al trabajar con Árbol de regresión, se presentan a continuación:

Tabla 15: Predicción de aperturas y criterios de evaluación por segmento con Árbol de regresión.
Fuente: Elaboración propia.

Segmento	Mínimo	Máximo	R^2	MAE	MAPE	RMSE
Segmento 1	1,37	2,27	0,02	1,54	2,66	1,95
Segmento 2	0,00	0,00	0,00	0,00	Inf.	0,02
Segmento 3	6,76	18,81	0,71	1,48	0,11	1,82
Segmento 4	1,83	5,52	0,17	2,43	Inf.	3,06
Segmento 5	9,16	9,16	0,00	2,13	0,26	2,61

Las **Tablas 14** y **15** dejan ver que el segmento 2 es el grupo para el que se predice, con ambos modelos, un menor intervalo de correos que se abren. Más aún, la cantidad de aperturas predichas para este grupo es prácticamente 0. Este resultado tiene sentido si consideramos que se trata de un *cluster* formado por aquellas personas saturadas que poseen un *open rate* muy cercano a 0% en promedio, es decir, está compuesto por los individuos que menos correos leen. Por ende, la cantidad de correos a enviar a estos individuos debería ser cercana a esa cifra.

Las tablas muestran además que el mayor intervalo de correos abiertos predicho por los dos modelos, corresponde al rango de aperturas del tercer *cluster*, lo cual concuerda con que este grupo está formado por los "mejores lectores" de correos, en otras palabras, los que más mails abren en promedio. De esta forma, las personas del segmento 3 debiesen ser quienes más correos reciban mes a mes.

Respecto a los criterios de evaluación, estos vuelven a evidenciar que, para la mayoría de segmentos, los modelos y las predicciones tienen un bajo desempeño. Si bien los segmentos 1, 2 y 5 no presentan las peores métricas de error, son los *clusters* en que los modelos menos se ajustan a sus datos, según el R^2 resultante para estos grupos en ambas técnicas. Por otro lado, el segmento 4 presenta una mejora del R^2, pero destaca por ser el *cluster* con las peores métricas de error.

A pesar del bajo desempeño que muestran ambos modelos para la mayoría de los segmentos, es importante resaltar que, para el tercer *cluster*, ambos modelos tienen un alto ajuste si se miran las cifras del R^2. Más aún, para este grupo, las métricas MAE, MAPE y RMSE suelen ser las más bajas y no están indeterminadas. De esta forma, es válido concluir que este segmento posee registros más homogéneos y por ende, las predicciones están más acertadas.

Finalmente, dados los resultados expuestos en esta sección, se define el número de correos a enviar por segmento mediante el modelo de Regresión lineal múltiple por sobre el Árbol de regresión. A pesar de que este segundo método presente predicciones con rangos de aperturas más acotados y sin mínimos negativos, se decide trabajar con el primer método mencionado, ya que suele presentar un ajuste mayor a los datos y un mejor desempeño de sus pronósticos. Además, se da una nueva interpretación a los intervalos de apertura predichos, donde se le otorga una mayor importancia a la cota máxima de estos rangos, y así establecer límites de correos a enviar por persona según el *cluster* al que pertenezcan. Con esto, los resultados para el número de mails a mandar por persona se presentan en la sección "Despliegue: Cantidad de correos a enviar"

7.4.3 Horario de envío

En esta parte se explica la modelación con la que se pretende encontrar un día y una hora en que sea más conveniente enviar los correos. En otras palabras, la finalidad de utilizar los modelos de Regresión logística binaria y Árbol de clasificación, es encontrar un horario de envío de mails que maximice una predicción de la probabilidad de lectura de estos mismos.

Por esto, se hace uso de los datos presentes en la Base de correos, los cuales son particionados según la clase de correo, tal como se muestra en el **Anexo A**. Para las observaciones del ámbito Fidelizar y tipo Tarjeta, es necesario obtener una muestra mediante la técnica de Muestreo aleatorio simple, ya que la capacidad de procesamiento de *RStudio* no permite incorporar los casi 40 millones de registros que presenta esta clase de correos. Así, la muestra que se utiliza en los modelos para este caso, contiene cerca de 10 millones de observaciones, y es obtenida procurando que la distribución de correos abiertos, días de envío y horas de envío, no se alejen en más de 5 p.p. respecto a la disposición poblacional correspondiente.

Una vez particionados los datos tal como se menciona anteriormente, se definen las variables a utilizar en los modelos. Como se trabaja con predicciones de la probabilidad de apertura de un correo, se utiliza como variable dependiente el índice binario que indica si un mail fue leído o no. Mientras que las variables independientes corresponden al día y la hora de envío del correo respectivo, de forma tal que así se puede conocer el efecto que tienen los horarios en que se manda un mail, sobre las lectura del mismo.

Los modelos de Regresión logística binaria y Árbol de clasificación, en su mayoría, son entrenados con un 70% de los registros de la clase de correos que se esté modelando. Así, las observaciones utilizadas para probar los modelos y predecir las probabilidades de apertura según el horario de envío, son el 30% de los datos restantes, correspondientemente. Además, los registros son divididos en su mayoría respecto a la variable Abierto. Esta sería la repartición utilizada por defecto, sin embargo, existen clases de correos que sus datos no pueden ser probados en los modelos dado este reparto, por lo que se modifican los parámetros para dividir los registros tal como se puede ver en el **Anexo F**.

De esta forma, se predice la probabilidad de apertura según el día y la hora de envío para las 20 clases de mails resultantes de la combinación ámbito y tipo de correo, sin aún especificar el segmento de personas. Así, los pronósticos tienen el siguiente formato, utilizando como ejemplo los correos del ámbito Pyme y tipo Pyme, a nivel general:

Tabla 16: Predicción de probabilidad de apertura según horario de envío para correos del ámbito Pyme y tipo Pyme en Regresión logística binaria.
Fuente: Elaboración propia.

Hora/Día	**Lunes**	**Martes**	**Miércoles**	**Jueves**	**Viernes**
7	0,280	0,272	0,267	0,257	0,260
8	0,305	0,297	0,292	0,281	0,284
9	0,286	0,278	0,273	0,262	0,266
10	0,318	0,310	0,305	0,293	
12	0,294	0,286	0,281	0,270	0,274
13	0,307	0,299	0,294	0,283	0,286
14				1,000	
15			0,359	0,347	
22	0,143				

Tabla 17: Predicción de probabilidad de apertura según horario de envío para correos del ámbito Pyme y tipo Pyme en Árbol de clasificación.
Fuente: Elaboración propia.

Hora/Día	**Lunes**	**Martes**	**Miércoles**	**Jueves**	**Viernes**
7	0,313	0,294	0,261	0,253	0,240
8	0,230	0,300	0,308	0,272	0,287
9	0,250	0,271	0,167	0,271	0,333
10	1,000	0,400	0,317	0,281	
12	0,260	0,263	0,286	0,261	0,327
13	0,284	0,305	0,269	0,296	0,275
14				1,000	
15			0,317	1,000	
22	0,143				

Las **Tablas 16** y **17** muestran el formato que tienen los resultados para las predicciones de las probabilidades de apertura de correos, según el horario de envío del mail. Estas tablas presentan en sus ejes verticales los niveles para la variable Hora de envío, mientras que en el eje horizontal están los valores que puede tener la variable Día de envío. Así, una combinación entre día y hora, especifica el valor predicho para la probabilidad de apertura de un correo perteneciente a una clase en particular, en el respectivo horario. Con esto, la resta entre 1 y el valor de alguna de estas predicciones, corresponde a la probabilidad de éxito de la no apertura de un correo, según el horario de envío del mail.

Es importante mencionar que los espacios vacíos de las tablas (marcados con color plomo) corresponden a horarios de envío de correos que el banco no suele utilizar, o no ha usado, para mandar los mails de una determinada clase. Por ende, estas combinaciones de día y hora de envío, no aparecen entre los datos que se implementan para probar los modelos y tampoco permiten predecir una probabilidad de apertura.

Las tablas expuestas y el criterio de evaluación usado en esta parte, presentan resultados que permiten discutir sobre el modelo a utilizar para diseñar las políticas de toques. Por un lado, el Árbol de clasificación se ajusta levemente mejor a los datos según lo que evidencia su *residual deviance*, la cual es entre 0,01% y 0,8% menor que la resultante para la Regresión logística binaria, si se revisa este parámetro en los modelos implementados para todos los conjuntos de datos que da la partición según la clase de correo.

No obstante, con el Árbol de decisión, se dificulta la selección de un horario de envío que maximice la probabilidad de apertura cuando existen paridades entre las predicciones. Un caso de esto puede verse en la **Tabla 17**, donde existen 3 horarios que predicen una probabilidad de apertura máxima: lunes entre 10:00:00 hrs. y 10:59:59 hrs., jueves entre 14:00:00 hrs. y 14:59:59 hrs. y jueves entre 15:00:00 hrs. y 15:59:59 hrs.

En cambio, la Regresión logística binaria permite definir el horario de envío de acuerdo a la significancia estadística[26] que presenten los estimadores del efecto de los días y horas en que se manda un correo, sobre el *odd ratio* de apertura. De esta forma, a través de este modelo, los horarios pueden ser escogidos de acuerdo los niveles de las variables Día de envío y Hora de envío que evidencien una mejora significativa en este factor. Así, con la predicción mostrada en la **Tabla 16**, el horario elegido para realizar los envíos de mails del ámbito Pyme y tipo Pyme sería el lunes entre 13:00:00 hrs. y 13:59:59 hrs., ya que el resto de horarios que predicen una probabilidad de apertura mayor, se caracteriza porque el nivel del día y/o la hora de envío, no evidencian un efecto significativo sobre el *odd ratio* de lectura de un correo. Por ende, en casos que ninguno de los niveles en una de las variables independientes presente una mejora significativa en el *odd ratio*, según el modelo de Regresión logística binaria, el horario para mandar correos se escoge según la significancia estadística que tengan los niveles de la otra variable independiente sobre este parámetro.

Dado lo anterior, los días y horas de envío de correos se definen según las predicciones obtenidas con el modelo de Regresión logística binaria Con esto, se verifican ciertos resultados en que el horario que pronostica una probabilidad de apertura máxima, parecería no ser el mejor para enviar los mails. En la tabla a continuación es posible ver un caso de esto para los correos del ámbito Vender y tipo Inversiones:

Tabla 18: Predicción de probabilidad de apertura según horario de envío para correos del ámbito Vender y tipo Inversiones en Regresión logística binaria.
Fuente: Elaboración propia.

Hora/Día	**Lunes**	**Martes**	**Miércoles**	**Jueves**	**Viernes**
6		0,444			
8	0,296	0,254	0,225	0,252	0,271
9	0,275	0,235	0,207	0,233	0,251
10	0,257	0,218	0,192	0,216	0,234
11	0,320	0,275	0,245		
12	0,297	0,254	0,225	0,252	0,272
13	0,340	0,294	0,262	0,292	
14	0,327	0,282	0,251	0,280	
15	0,302	0,259	0,230	0,257	
16	0,283	0,242	0,214		

La **Tabla 18** muestra que para esta clase de correos, el horario de envío que predice una mayor probabilidad de apertura es el martes entre 6:00:00 hrs. y 6:59:59 hrs. (marcado en color naranjo). Además, estos niveles para el día y hora de envío respectivamente, presentan una mejora estadísticamente significativa en el *odd ratio* de la apertura de correos, por lo que sería un horario idóneo para mandar los mails. Sin embargo, el modelo de Regresión logística binaria muestra que el día lunes tiene un mejor efecto significativo sobre este parámetro si se compara con el día martes. No obstante, se evidencia que el horario del lunes entre 6:00:00 hrs. y 6:59:59 hrs. (marcado en color rojo), no posee registros en el conjunto de datos utilizado en este caso, dando a entender que

[26] Se utiliza un nivel de significancia del 10%.

no ha sido utilizado por el banco para realizar el envío de mails de la clase Vender Inversiones, por lo que se propone y queda a criterio de la organización si comienza a ser implementado o no.

Finalmente, conforme lo explicado a lo largo de esta esta sección, se realizan los pronósticos de la probabilidad de apertura de un correo según el horario de envío del mismo, para cada una de las 20 clases en estudio y considerando el *cluster* al que pertenecen las personas que recibieron los mails. El segmento de clientes es incorporado en la modelación utilizando los registros de correos que han sido enviados a personas de ese grupo en particular. Así, por ejemplo, para predecir la probabilidad de apertura de los correos del ámbito Vender y tipo Consumo en el tercer segmento de clientes, se seleccionan las observaciones correspondientes a esa clase de correos y que fueron enviados a personas de ese *cluster.* De esta forma, siguiendo un procedimiento análogo al utilizado para los casos en que no se especifica el segmento de la persona contactada, se obtienen los resultados expuestos en la sección "Despliegue: Horarios de envío" y en el **Anexo G**.

7.5 Despliegue

En esta parte se muestran los resultados finales del trabajo, los que se obtienen luego de incorporar los análisis que se llevan a cabo en la sección "Desarrollo metodológico: Modelación y evaluación".

7.5.1 Segmentación

Luego de realizar la segmentación mediante Agrupamiento jerárquico aglomerativo, a una de las muestras de la Base de personas, este *clustering* se utiliza para entrenar el modelo de *K*-vecinos más cercanos que permite clasificar a todos los individuos de la base en cuestión, en los segmentos obtenidos para la muestra. De esta forma, utilizando $k = 1$, es decir, que el vecino más próximo a una "nueva" persona determine el *cluster* al que pertenece esta última, se obtiene la siguiente segmentación para los 405.312 individuos en estudio:

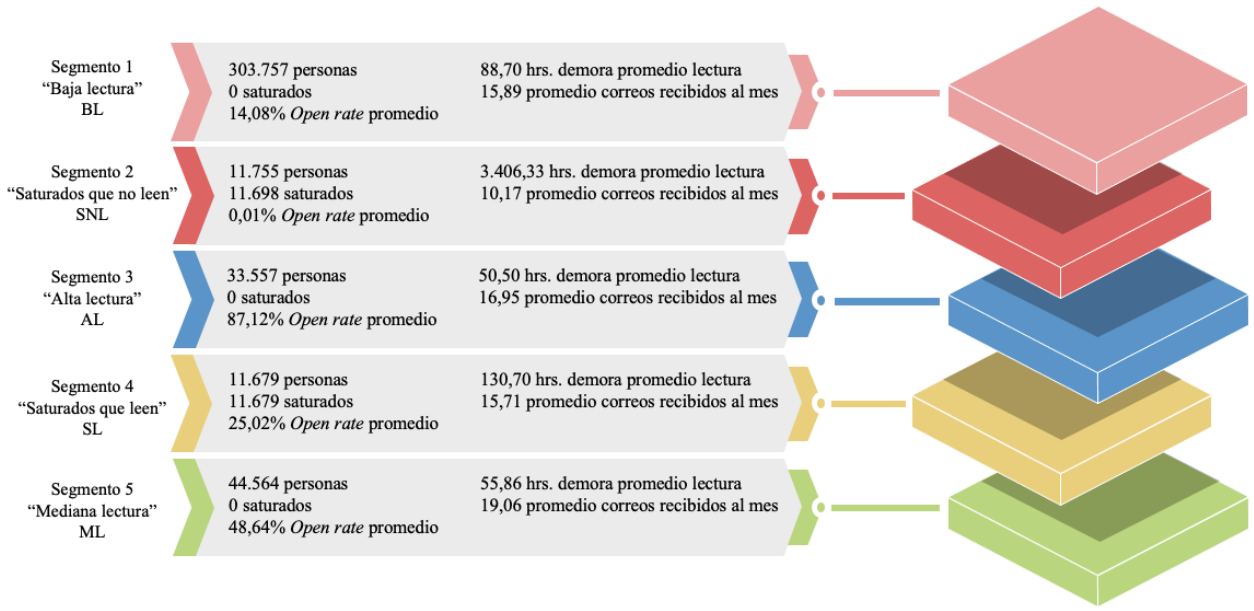

Figura 21: Segmentación para Base de personas.
Fuente: Elaboración propia.

La **Figura 21** muestra diferentes características con las que cuentan los segmentos resultantes. Entre ellas, destaca que existe un segmento, denominado de "Baja lectura" (BL), que agrupa a la mayoría de las personas, concentrando casi un 75% de la población en estudio. Este *cluster* se reconoce por ser el grupo que, en promedio, abre menos correos y que demora más en leer los mismos, de entre los 3 conjuntos que no incorporan individuos no saturados. Siguiendo esta línea de los grupos de personas no saturadas, existe uno, nombrado como "Alta lectura" (AL), que su cualidad principal es la elevada tasa de apertura de correos que promedia, lo que se condice con que sea el segmento de menor demora de lectura media. Además, está el grupo de "Mediana lectura" (ML) que promedia un *open rate* cercano al 50% y resalta por ser el *cluster* que recibe la media de correos más alta. Llama la atención que este último aspecto no sea una característica del segmento AL, para así aprovechar el alto nivel de lectura que muestra este último conjunto.

Por otro lado, existen dos segmentos que están conformados completamente o en su mayoría por individuos saturados: el grupo "Saturados que no leen" (SNL) y el conjunto "Saturados que leen" (SL), respectivamente. El primero, se caracteriza por promediar una tasa de lectura de correos de casi 0%, lo que refleja una demora media de lectura de más de 3 mil horas. Esto evidencia la poca receptividad que este grupo tiene para con el email marketing del banco, dando sentido a que sean las personas que menos mails reciban al mes, en promedio.

El otro conjunto de personas saturadas lee casi 1 de cada 4 correos que recibe, reflejando una tasa media de lectura mejor que el segmento BL, por lo que contrario a lo que evidencia el promedio de envíos, este grupo de individuos saturados podría ser contactado en una proporción mayor al conjunto de no saturados de menor tasa de apertura.

Revisando las variables del *clustering* que hacen referencia a la cantidad de productos, se obtienen los resultados mostrados a continuación para el número promedio de estos por persona, en cada segmento:

Tabla 19: Cantidad de productos promedio por persona según segmento de clientes.
Fuente: Elaboración propia.

Segmento	Tarjeta de crédito	Inversiones	Seguros	Cuenta prima	Crédito de consumo	Crédito hipotecario
BL	1,30	1,01	1,65	0,91	0,54	0,21
SNL	1,28	1,11	1,33	0,73	0,42	0,18
AL	1,44	1,55	1,84	0,89	0,55	0,20
SL	1,23	0,90	1,62	0,93	0,56	0,19
ML	1,48	1,60	1,81	1,08	0,61	0,28

La **Tabla 19** refleja que el segmento ML es aquel que generalmente posee un mayor número de productos promedio por persona, lo cual se evidencia en las cifras para tarjeta de crédito, inversiones, cuenta prima, crédito de consumo y crédito hipotecario. Lo anterior podría explicar el hecho de que este *cluster* sea el que reciba una media mensual de correos mayor. Siguiendo esta línea, el segmento AL suele aparecer por detrás del grupo de mediana lectura al revisar este aspecto de la cantidad promedio de artículos por persona. Ante esto, no es erróneo concluir que se evidencian mayores niveles de lectura mientras más alta sea la cantidad media de productos que registren los individuos de un segmento.

Mediante la tabla expuesta, también se verifica que los segmentos de menores ratios de apertura de correos, se caracterizan por contar con las cantidades de productos más bajas por persona en promedio. Más aún, por lo general, uno de los segmentos de individuos saturados aparece con la menor cifra del número promedio de un artículo por persona, para los 6 productos en cuestión: SNL en seguros, cuenta prima, crédito de consumo y crédito hipotecario; mientras que SL para tarjeta de crédito e inversiones. Esto podría explicar el hecho de que las personas saturadas sean las que, en promedio, menos correos reciban mes a mes por parte del banco. Además, lo anteriormente detallado, estaría reflejando un menor grado de fidelidad con la organización por parte los individuos que conforman los dos *clusters* mencionados.

Cabe resaltar que aquellas variables de la Base de personas, que no se mencionan en esta parte, no evidencian mayores diferencias si se comparan entre segmentos y se aproximan a las proporciones que presenta la base sin segmentar.

7.5.2 Cantidad de correos a enviar

Tal como se menciona hacia el final de la sección “Desarrollo metodológico: Modelación y evaluación – Cantidad de correos a enviar”, en esta parte se interpreta de una forma distinta las predicciones resultantes. Con esto, el foco se posa sobre el valor máximo que entregan los

intervalos predichos mediante Regresión lineal múltiple, los cuales se aproximan por exceso a la unidad, de forma tal que se establece un límite de toques a realizar por mes a través de email, según el segmento de la persona contactada. Estas cifras pueden verse en la tabla a continuación:

Tabla 20: Límite de correos a enviar por mes según segmento.
Fuente: Elaboración propia.

Segmento	Nº correos
BL	4
SNL	1
AL	22
SL	8
ML	11

La **Tabla 20** presenta el número de correos que como mucho puede recibir un cliente mensualmente, de acuerdo al *cluster* en que este mismo esté asignado. Así, es posible ver que las personas del segmento de alta lectura, serían las que más correos podrían recibir mes a mes, al evidenciarse que el límite de este grupo se establece en 22 mails. No es raro obtener un resultado de este estilo considerando que el *cluster* AL se caracteriza principalmente por su elevado *open rate* promedio, lo cual se vincula con mayores dígitos en las predicciones de aperturas de correos. Así, se espera que los individuos de este segmento sean los principales lectores del email marketing de la organización.

Revisando el resto de resultados expuestos en la tabla, se aprecia que, por detrás del segmento AL, aparecen los grupos de mediana lectura y de personas saturadas que leen correos, con límites de 11 y 8 mails que podrían recibir cada individuo de estos *clusters* por mes, respectivamente. Estas cifras buscan que las personas que conforman los grupos ML y SL, tengan un aumento gradual en sus niveles de apertura, para que así puedan ser incluidos en el segmento de alta lectura.

Los segmentos BL y SNL aparecen con las cifras más bajas en la tabla, es decir, las personas de estos grupos recibirían una cantidad menor de correos mensualmente. Para el *cluster* de baja lectura, se espera que el límite propuesto ayude a disminuir los niveles de desuscripción que tiene el email marketing del banco. En el caso de las personas saturadas que no leen, si bien el intervalo predicho para la cantidad de correos a enviar a estos individuos, es bastante cercano a cero, se propone el límite de 1 correo mensual para que así la empresa pueda mantener el contacto por email con estos clientes, y evitar que salgan de la base de contactos con la que trabaja el banco.

Finalmente, cabe mencionar que estos resultados modificarían directamente el parámetro "Saturación" de la política de saturación que utiliza el banco, que plantea que no se pueden mandar más de 15 correos al mes por persona, independiente de quien sea contactado. No obstante, los resultados de la **Tabla 20** pueden utilizarse con las otras condiciones que establece este parámetro: enviar máximo 1 correo diario y a lo más 7 por semana, para cada individuo.

7.5.3 Horarios de envío

Tal como se detalla en la sección "Desarrollo metodológico: Modelación y evaluación – Horarios de envío", los segmentos de clientes se incorporan en los modelos de Regresión logística binaria para las 20 clases de correos con las que se trabaja. De esta forma, los horarios para mandar mails, según las predicciones de probabilidad de apertura de correos, quedan de la siguiente manera, tomando como referencia nuevamente los registros del ámbito Vender y tipo Inversiones:

Tabla 21: Horarios de envío de correos del ámbito Vender y tipo Inversiones.
Fuente: Elaboración propia.

	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	6:00:00 - 6:59:59	Lunes	6:00:00 - 6:59:59
BL	Martes	6:00:00 - 6:59:59	Viernes	6:00:00 - 6:59:59
SNL	Miércoles	15:00:00 - 15:59:59	-	-
AL	Martes	6:00:00 - 6:59:59	-	-
SL	Viernes	8:00:00 - 8:59:59	-	-
ML	Lunes	11:00:00 - 11:59:59	Lunes	6:00:00 - 6:59:59

La tabla anterior muestra, en su columna "Recomendado", el día y hora de envío que se sugiere para que el banco mande los mails de esta clase de correos, según el segmento que integre la persona contactada. También se incluye, en la fila "General", el horario recomendado cuando no se conoce el *cluster* al que pertenece un individuo. Estos resultados se caracterizan por haber sido usados por la empresa en el email marketing de esta clase de correos. Por otro lado, en la columna "Propuesto", aparecen los horarios que parecerían tener un mejor efecto en la probabilidad de apertura de los correos si se comparan con los recomendados, pero que no suelen ser usados por la organización para realizar los envíos, de forma tal que estas combinaciones de día y hora aparecen en menor medida entre los datos de envío de correos. Por ende, se deja a criterio del banco si estos horarios son implementados o no.

Todos los resultados de esta parte, para las 20 clases de correos estudiadas, se presentan en el **Anexo G**. Allí es posible ver que hay clases donde, entre los horarios recomendados para contactar a los diferentes segmentos, predomina un día o existe intervalo más acotado de horas de envío. Lo primero se refleja, por ejemplo, en los mails del ámbito Vender y tipo Aumento cupo, los cuales en su mayoría debiesen ser enviados los viernes. El segundo aspecto se puede observar en la clase Fidelizar PAT, que presenta un rango de 3 horas para realizar los envíos. No obstante, existen casos en que ocurre lo contrario, es decir, no existe un día o una intervalo de horas predilecto para mandar los mails de una misma clase. Por ejemplo, para el envío de correos del ámbito Vender y tipo Cuotización, los horarios presentan 4 niveles de días para contactar a los distintos segmentos, como también cuentan con un rango de 12 horas. Esto demuestra la homogeneidad y heterogeneidad que pueden tener los niveles de días y horas, entre los horarios recomendados para contactar a los segmentos de personas, de acuerdo a la clase de correos en consideración.

Otro aspecto a destacar es que, recordando lo expuesto en la **Tabla 8,** el miércoles es el segundo día en que más correos se envían por parte del banco. Sin embargo, los resultados de esta parte muestran que este es el día hábil que menos se repite entre los horarios recomendados para contactar a los diferentes segmentos, según la clase del correo enviado, lo cual quitaría validez para que el miércoles siga siendo uno de los días hábiles más utilizados por el banco para mandar mails.

En cuanto a las horas de envío, destaca que el intervalo entre las 14:00:00 hrs. y 14:59:59 hrs. sea el rango que más se repite entre los horarios recomendados. Esto toma mayor relevancia si se considera lo que muestra el **Anexo B**, donde este intervalo es el segundo más utilizando por el banco para mandar correos. Situación similar ocurre con el rango 12:00:00 hrs. a 12:59:59 hrs., que es el más utilizado para enviar mails, y también es de las horas que más frecuentan los horarios recomendados. En esta línea, también es importante comentar lo que ocurre con intervalos de horas que en el **Anexo B** evidencian altas tasas de lectura como los rangos 01:00:00 hrs. a 01:59:59 hrs., y 18:00:00 hrs. a 18:59:59 hrs., los cuales no aparecen mucho entre los horarios recomendados, reflejando que no serían horas con buenas aperturas como se plantea en el anexo mencionado.

Tal como se menciona en la sección "Desarrollo metodológico: Preparación de los datos – Preparación bases de datos", el área del CRM Personas del banco, buscando una mayor lectura del email marketing por parte de los clientes, generalmente envía los correos de lunes a viernes entre las 7:00 hrs. y las 19:00 hrs. aproximadamente. Aún así, llaman la atención unos casos, según la clase del correo y el segmento del individuo contactado, en que la predicción para la probabilidad de apertura del mail, se maximiza en un horario de envío que esta fuera de este rango mencionado, lo que se contradice con ciertas hipótesis planteadas dentro de la empresa, las cuales proponen que el envío de correos los fines de semana y/o a horas de la madrugada o la noche, disminuyen el nivel de apertura de estos mismos.

Revisando los resultados para los principales ámbitos, es posible verificar que para el caso de Fidelizar, los horarios recomendados suelen estar compuestos por los días lunes o jueves, y existe una predominancia aún más clara de las horas pm. Más aún, este último aspecto se refleja claramente en 4 de los 6 tipos de correos que presenta este ámbito. Siguiendo con Informar, existe una leve diferencia que favorece a las horas am, sin embargo, lo que más resalta es que los horarios recomendados para este ámbito, están conformados en su mayoría por el viernes, y más destacable es el domingo, día muy poco utilizado por el banco para enviar correos, tal como se muestra en la **Tabla 8**. En el caso de Ventas, sobresale que el martes sea el nivel para el día de envío que aparece con mayor frecuencia. Por añadidura, este día define la mayoría de los horarios recomendados para 4 de los 8 tipos de correos con los que cuenta este ámbito. Respecto a las horas de envío en este caso, no existe mayor predominancia de alguna jornada.

Análogo a lo comentado anteriormente, pero esta vez para los tipos de correo más enviados por el banco, destaca que en el caso de Tarjeta, existen dos intervalos de horas para realizar los envíos de mails, según el ámbito del mismo: durante la mañana o medio día para aquellos correos que tratan de vender, mientras que los mails que buscan fidelizar, más adentrados en la tarde e incluso cerca del fin de la jornada laboral diurna. Para Canales digitales, se tiene que los horarios recomendados están formados en su mayoría por el día domingo y un rango de pocas horas, con lo que podría incluso usarse solo un horario para enviar estos correos a todas las personas. En cuanto al tipo

Otros, se verifica que el envío de mails debiese realizarse mayormente en las primeras horas de la mañana, independiente del segmento contactado y el día de envío.

Revisando los resultados de esta parte según los segmentos de personas, es posible verificar que el martes es el día que suele predominar entre los horarios recomendados para contactar a individuos de los *clusters* AL, BL y ML, o sea, personas no saturadas. Esto toma aún más relevancia al recordar que estos grupos son los que más gente incorporan, lo que podría significar en una brecha aún mayor que la evidenciada en la **Tabla 8** a favor del martes. Contrario a esto es lo que ocurre con los segmentos de personas saturadas SL y SNL, los cuales presentan horarios conformados en su mayoría por los días viernes y lunes respectivamente. En cuando a las horas de envío, nuevamente es posible diferenciar según la saturación de las personas: los *clusters* AL, ML y BL presentan una mayor frecuencia de sus horarios en el intervalo 14:00:00 hrs. a 14:59:59 hrs., aunque el segmento de baja lectura posee una segunda mayoría de sus horarios en el rango 7:00:00 hrs. a 7:59:59 hrs., mientras que la hora que más se repite entre los resultados para los grupos SN y SNL, corresponde al intervalo 8:00:00 hrs. a 8:59:59 hrs.

En cuanto a los horarios propuestos, se obtienen 23 de estos resultados. En su mayoría, estos horarios se caracterizan por cambiar el día o la hora que se especifica en los horarios recomendados de envío. Más aún, solo uno de los propuestos cambia ambos componentes del recomendado.

Así, se tiene que más de la mitad de estos horarios propuestos plantean modificar el envío de correos para que se realicen el día lunes. Siguiendo esta línea, destaca que más de un cuarto propone cambiar el día de un horario recomendado por el domingo. Además, una proporción similar se refleja entre los horarios propuestos que ofrecen salir de la jornada que suele utilizar el banco, y realizar los envíos antes de las 7 am. Estos dos últimos aspectos evidencian nuevamente que los horarios que no están dentro de la jornada que establece internamente el banco para realizar los envíos, parecerían no disminuir el grado de recepción del email marketing tal como cree la organización.

8. Diseño experimental

Según lo expuesto en las secciones anteriores, existe evidencia de que en el email marketing del banco se pueden utilizar diferentes políticas de toques con el fin de mejorar las tasas de apertura de los correos. Ante esto, los resultados obtenidos en este trabajo deben ser complementados con experimentos, para así confirmar si estos hallazgos tienen efectos reales sobre el nivel de recepción que tengan las personas de las campañas de *emailing* que implementa la organización. A continuación se detalla un diseño experimental que permitiría probar lo mencionado.

8.1 Hipótesis por probar

- Existe un límite de correos a enviar por segmento: Por cada segmento, existe un máximo de correos que las personas están dispuestas a leer.
- Existen horarios en que es más probable que se abra un correo: Por cada segmento y clase de correo, existe un día y una hora en que es más conveniente enviar los correos dado que es más probable que el mail sea leído.
- Existen componentes del asunto de los correos que tienen un efecto positivo en la apertura[27]:
    - Modo imperativo.
    - Exclamaciones en vez de preguntas.
    - Beneficio explícito.
    - Palabras claves:
        - "Importante".
        - "Felicitaciones".

8.2 Variables experimentales

- Dependiente:
    - Apertura de correos: Variable que indica la cantidad de correos que se leen. Se puede considerar como una tasa continua en el rango [0,100] o como una binaria.
- Independientes:
    - Cantidad de envíos mensuales: Se estudia el efecto de exceder un límite de envíos mensuales sobre la cantidad de aperturas. Esta variable permitiría confirmar si existe un efecto de saturación en la lectura, es decir, las personas dejan de leer correos frente a un exceso de envíos.
    - Horario de envío: Se estudia el efecto que tiene el día y la hora en que se envía un correo, sobre la apertura del mismo. De esta forma se podría determinar qué horario es el más conveniente para enviar un mail.
    - Segmento: Corresponde al segmento de clientes en que se clasificó a una persona respecto a su nivel de lectura. Según esta variable, varía la cantidad de correos a mandar y los horarios de envío a utilizar.
    - Clase del correo: Variable que indica el ámbito y el tipo de correo en estudio.
    - Asunto: Se estudia el efecto que tienen diferentes formatos del asunto de los correos. Estos formatos podrían incluir los siguientes aspectos:
        - Modo imperativo: El asunto expresa una orden o mandato.
        - Exclamaciones en vez de preguntas: El asunto utiliza exclamaciones por sobre preguntas abiertas.
        - Beneficio explícito: Se manifiesta un beneficio ofrecido por la empresa en el asunto.
        - Palabras clave: El asunto incluye las palabras "Importante" o "Felicitaciones".

[27] Si bien no está dentro de los alcances de la memoria, se incluyen en el diseño experimental por petición de la organización.

8.3 Muestra de clientes

La muestra inicial de clientes corresponde a aquellos que se han considerado para realizar este trabajo de título: 405.312 personas cuentacorrentistas que han sido contactadas por email todos los meses entre julio del 2020 y octubre del 2021, excepto febrero del 2021. Sin embargo, este número queda sujeto al conjunto de clientes que la empresa quiera utilizar para realizar la experimentación, y a la actualización de la muestra al incluir datos más recientes del email marketing del banco.

8.4 Grupo de control

Es importante que la muestra de clientes sea dividida de manera aleatoria en dos grupos: uno de tratamiento, al cual se le aplicarán las condiciones de email marketing a experimentar; y uno de control, el cual no recibirá tratamiento y se contactará mediante la política de toques tradicional. Se propone que esta división se realice a través de Muestreo aleatorio simple, con el que se espera obtener dos grupos con cerca del 50% de individuos que posee la muestra inicial de clientes. Es importante que ambos grupos tengan distribuciones similares a las que tiene la muestra inicial, en variables demográficas como edad, género, nivel educacional y estado civil, para que de esta forma los grupos sean homogéneos y los resultados de la experimentación estén determinados por el efecto del tratamiento y no por la composición de ambos grupos. En otras palabras, al utilizar este mecanismo, se pretende reducir el sesgo que pudiese tener el experimento.

8.5 Experimentos

Para llevar a cabo la experimentación, es necesario incorporar ciertas modificaciones al email marketing que realiza la empresa, para así poder medir el tratamiento en el contexto natural de cada individuo. Sin embargo, este tratamiento estará expuesto a diferentes externalidades que podrían generar ruido o sesgo en la medición, por ejemplo, que una persona no lea su correo por un motivo de fuerza mayor. A pesar de ello, la experimentación debe realizarse dentro de este contexto natural dada la imposibilidad de medir el tratamiento a través un estudio más controlado.

Así, la experimentación considera que el email marketing que se debe utilizar para contactar a un individuo perteneciente al grupo de control, no tiene que modificarse respecto al que la empresa utiliza normalmente. Mientras que para comunicar a una persona del grupo tratamiento, el *emailing* debe considerar cambios en las variables experimentales independientes presentadas anteriormente. Con esto, se plantean 3 alternativas de experimentos, las cuales se detallan a continuación:

- Experimentar sobre cantidad de envíos: Estos experimentos permiten verificar la primera hipótesis, la cual plantea que existen diferentes límites para el número de envíos que se le pueden realizar a los individuos de cada segmento. De esta forma, para aquellas personas que pertenezcan al grupo tratamiento, debe respetarse esta cantidad máxima de envíos según su segmento, mientras que aquellos individuos que estén en el grupo control, la cantidad de mails

que reciban tiene que ser mayor al límite planteado. A continuación, se proponen los siguientes experimentos de este estilo para llevar a cabo:

Tabla 22: Propuesta de experimentos según cantidad de envíos.
Fuente: Elaboración propia.

Hipótesis	Grupo tratamiento	Escenario tratamiento	Grupo control	Escenario control
Las personas del segmento AL leen a lo más 4 correos en un mes.	50% de personas pertenecientes al segmento AL.	Se envían 4 correos en un mes.	50% de personas restantes pertenecientes al segmento AL.	Se envían más de 4 correos en un mes.
Las personas del segmento SNL leen a lo más 1 correo en un mes.	50% de personas pertenecientes al segmento SNL.	Se envía 1 correo en un mes.	50% de personas restantes pertenecientes al segmento SNL.	Se envían más de 1 correos en un mes.
Las personas del segmento BL leen a lo más 22 correos en un mes.	50% de personas pertenecientes al segmento BL.	Se envían 22 correos en un mes.	50% de personas restantes pertenecientes al segmento BL.	Se envían más de 22 correos en un mes.
Las personas del segmento SL leen a lo más 8 correos en un mes.	50% de personas pertenecientes al segmento SL.	Se envían 8 correos en un mes.	50% de personas restantes pertenecientes al segmento SL.	Se envían más de 8 correos en un mes.
Las personas del segmento ML leen a lo más 11 correos en un mes.	50% de personas pertenecientes al segmento ML.	Se envían 11 correos en un mes.	50% de personas restantes pertenecientes al segmento ML.	Se envían más de 11 correos en un mes.

- Experimentar sobre el día y la hora de envío: Estos experimentos permiten confirmar la segunda hipótesis, la que plantea que existen horarios en que es más probable que las personas lean los correos, según el segmento que integren y la clase de correo que se envía. Así, aquellos individuos que pertenezcan al grupo tratamiento, deben ser contactados un día y/o a una hora en particular, mientras que las personas del grupo control son contactadas según estipule la política de toques vigente. A continuación, se proponen experimentos de este tipo[28]:

[28] Estos experimentos pueden variar en segmento y clase de correo, aspectos que determinan el horario a testear y grupos a considerar, según los resultados obtenidos en este trabajo de memoria.

Tabla 23: Propuesta de experimentos según horario de envío.
Fuente: Elaboración propia.

Hipótesis	Grupo tratamiento	Escenario tratamiento	Grupo control	Escenario control
Las personas del segmento AL tienen más probabilidad de leer un correo de "Venta Inversiones" si se envía un martes entre las 6:00:00 hrs. y 6:59:59 hrs.	50% de personas pertenecientes al segmento AL.	Se envía un correo de "Venta Inversiones" un martes entre las 6:00:00 hrs. y 6:59:59 hrs.	50% de personas restantes pertenecientes al segmento AL.	Se envía un correo de "Venta Inversiones" en un horario distinto al martes entre las 6:00:00 hrs. y 6:59:59 hrs.
Las personas del segmento AL tienen más probabilidad de leer un correo de "Venta Inversiones" si se envía un martes.		Se envía un correo de "Venta Inversiones" un martes.		Se envía un correo de "Venta Inversiones" un día distinto al martes.
Las personas del segmento AL tienen más probabilidad de leer un correo de "Venta Inversiones" si se envía entre las 6:00:00 hrs. y 6:59:59 hrs.		Se envía un correo de "Venta Inversiones" entre las 6:00:00 hrs. y 6:59:59 hrs.		Se envía un correo de "Venta Inversiones" a una hora fuera del intervalo 6:00:00-6:59:59.

- Experimentar sobre el asunto del correo: Estos experimentos podrían verificar la tercera hipótesis, la cual plantea que existen componentes del asunto que favorecen la apertura de los correos. Aquellas personas que pertenezcan al grupo tratamiento, se les debe enviar un correo que en su asunto contenga alguna de las componentes mencionadas en la sección "Diseño experimental: Variables experimentales", mientras que a las personas del grupo control, se les manda un correo que en el asunto no cuente con ninguno de estos componentes. La siguiente tabla muestra experimentos de este estilo que se podrían aplicar[29]:

[29] Los experimentos propuestos en esta parte podrían modificarse según las campañas de email marketing que realice la empresa durante el periodo de experimentación. Sin embargo, los experimentos deben seguir esta lógica propuesta para sus asuntos.

Tabla 24: Propuesta de experimentos según asunto del correo.
Fuente: Elaboración propia.

Hipótesis	Grupo tratamiento	Escenario tratamiento	Grupo control	Escenario control
El uso del modo imperativo en el asunto de los correos aumenta la apertura.	50% de personas pertenecientes a la muestra inicial de clientes.	Se envía un correo que en su asunto ponga "*Paga en cuotas la deuda internacional de tu tarjeta de crédito*".	50% de personas restantes pertenecientes a la muestra inicial de clientes.	Se envía un correo que en su asunto ponga "*Recuerda que puedes pagar en cuotas la deuda internacional de tu tarjeta de crédito*".
Las exclamaciones en el asunto de los correos tienen un mejor efecto en la apertura que las preguntas abiertas.		Se envía un correo que en su asunto ponga "*Consejos para adquirir tu primera casa*".		Se envía un correo que en su asunto ponga "*¿Primera casa?, consejos para adquirirla*".
Mencionar un beneficio en el asunto de los correos, mejora la apertura.		Se envía un correo que en su asunto ponga "*Llévate este Tablet por contratar un seguro de auto*".		Se envía un correo que en su asunto ponga "*Llévate este beneficio por contratar un seguro de auto*".
Utilizar las palabras "importante" o "felicitaciones" aumenta la apertura de los correos.		Se envía un correo que en su asunto ponga "*Información importante sobre tu proceso hipotecario*".		Se envía un correo que en su asunto ponga "*Información sobre tu proceso hipotecario*".
		Se envía un correo que en su asunto ponga "*Felicitaciones, obtuviste $500.000 por comprar con las billeteras digitales y tu tarjeta*".		Se envía un correo que en su asunto ponga "*Obtuviste $500.000 por comprar con las billeteras digitales y tu tarjeta*".

8.6 Evaluación de experimentos

El primer experimento debe ser medido en un horizonte de tiempo mensual, es decir, debe ser implementado al menos un mes para obtener resultados. La medición de esta parte tiene que ver con el nivel de lectura que generen diferentes cantidades de correos enviados a lo largo de los distintos segmentos.

Así, para corroborar la hipótesis relacionada a este experimento, se debe utilizar una prueba de proporciones[6][36] y con esto verificar si existe una diferencia significativa en las tasas de apertura que tengan las personas, según los límites de envío propuestos. Con esto, la hipótesis nula (H_0) plantea que no existen diferencias entre el nivel de lectura del grupo tratamiento (π_T) y el del grupo control (π_C). Por el contrario, la hipótesis alternativa (H_1) plantea que efectivamente existen diferencias entre los niveles de apertura de estos grupos. La formulación de este método se presenta a continuación:

$$H_0: \pi_T = \pi_C \quad (20)$$

$$H_1: \pi_T \neq \pi_C \quad (21)$$

$$Z = \frac{\widehat{p_T} - \widehat{p_C}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_T} + \frac{\hat{p}(1-\hat{p})}{n_C}}} \sim N(0,1) \quad (22)$$

Con:

- Z: Estadístico de prueba.
- $\widehat{p_T}$: Estimador de la tasa de apertura en grupo tratamiento.
- $\widehat{p_C}$: Estimador de la tasa de apertura en grupo control.
- $\hat{p}$: Estimador de la tasa de apertura muestral.
- n_T: Número de observaciones en grupo tratamiento.
- n_C: Número de observaciones en grupo control.

En esta parte, se propone utilizar el promedio simple como estimador. De esta forma, si se cumple $Z > Z_{\alpha/2}$, se rechaza la hipótesis nula para un nivel de significancia α.

El segundo y tercer experimento se puede medir en un horizonte temporal de tres días por calendario, es decir, por ejemplo, si el tratamiento se implementa un lunes, se espera martes y miércoles para el jueves ver los resultados. Por ende, se entiende que la cantidad de aperturas no cambiará transcurridos dos días desde el envío[30].

[30] No obstante, este criterio puede cambiar considerando que los correos pueden ser abiertos en más o menos tiempo.

Para estos casos, los resultados estarán dados por el número de aperturas que se obtengan. De esta forma, para corroborar las hipótesis relacionadas a estas mediciones, se puede aplicar el método de diferencias en diferencias[32] y así medir el efecto tratamiento. La formulación de este se plantea a continuación:

$$\widehat{DD} = (\widehat{Y_1^T} - \widehat{Y_1^C}) - (\widehat{Y_0^T} - \widehat{Y_0^C}) \quad (23)$$

Con:

- $\widehat{DD}$: Estimador de diferencias en diferencias.
- $\widehat{Y_1^T}$: Estimador de aperturas para grupo tratamiento después del experimento.
- $\widehat{Y_1^C}$: Estimador de aperturas para grupo control después del experimento.
- $\widehat{Y_0^T}$: Estimador de aperturas para grupo tratamiento antes del experimento.
- $\widehat{Y_0^C}$: Estimador de aperturas para grupo control antes del experimento.

Nuevamente, se propone utilizar el promedio como estimador. Así, con el método de diferencias en diferencias, se puede estimar el efecto del tratamiento asumiendo que sin este, el resultado para los grupos serían similares.

9. Conclusiones

9.1 Conclusiones del trabajo

En primer lugar, este estudio permite evaluar la forma en que se realiza el email marketing en la empresa. Así, se verifica que existe un desajuste entre las políticas de toques que utiliza el banco y el grado de receptividad que las personas tienen para con los correos de la organización. Esto tiene un impacto importante para la empresa si se considera, por un lado, que el correo electrónico es de los principales canales por los que el banco se comunica con sus clientes, y por otro, los efectos negativos que se provocan sobre la venta de productos. De esta forma, se concluye que es importante generar políticas de email marketing personalizadas frente a la heterogeneidad que puedan presentar las personas en sus niveles de lectura de correos.

Ante la inexistencia de un parámetro para identificar personas saturadas por el email marketing del banco, se establece el criterio de "Caída en la tasa de apertura", con el que es posible considerar cerca de un 5% de los clientes en estudio como saturados, lo que permite concluir que este tipo de individuos se caracterizan por presentar disminuciones abruptas y prolongadas en sus niveles de lectura de correo, o bien, por haber leído un número menor de los correos del banco históricamente. Además, a través de este parámetro, se desprende que mientras menor sea el nivel de lectura inicial de una persona, mayor tiene que ser la caída, en proporción, para que este individuo sea considerado

saturado; mientras que las recuperaciones, en proporción, son similares para los diferentes intervalos de apertura inicial.

En cuanto al diseño de políticas de toques, es posible concluir que las personas que evidencian mejores niveles de lectura, son aquellas que el banco debiese aprovechar para hacer un mayor uso de su email marketing, dado que tienen un mayor grado de fidelidad con el banco, por lo que son capaces de leer cantidades más altas de mails. Por otro lado, para aquellos individuos que demuestran pocas aperturas de correos, tales como los clientes saturados o aquellos que mantienen una tasa de lectura de 0% mes a mes, la empresa debiese tener más cuidado en la cantidad de contactos que les realiza por mail y enfocar las comunicaciones por otro canal; ya que para estos individuos se hace más sencillo solicitar no ser contactados por email.

Además, se concluye que existen horarios en que es más conveniente contactar a las diferentes personas según la clase del correo a mandar. En otras palabras, hay días y horas de envío para los distintos mails que son más oportunos y ofrecen mayores posibilidades de lectura. Aquí es importante mencionar nuevamente que existen casos donde los horarios "idóneos" para mandar correos, se encuentran fuera del rango que suele utilizar la organización para mandar mails (de lunes a viernes entre 7:00 hrs. y 19:00 hrs.), permitiendo deducir que los envíos durante los fines de semana o en horas de la madrugada y la noche, parecerían no tener tan mala recepción como se cree, por lo que sería interesante contar con registros de estos horarios poco utilizados para evaluar esta última deducción.

A pesar de que en este trabajo existe un sesgo relacionado a las decisiones que se tomaron históricamente al momento de realizar los envíos de correos, se destaca que estos envíos tuvieron una parte aleatoria que se evidencia al no existir un único horario en que se llevaron a cabo, y al no contactar a solo una persona en particular.

Finalmente, es importante mencionar que, si bien los modelos llevados a cabo en este trabajo no se caracterizan por presentar buenos ajustes, los resultados obtenidos permiten dar una impresión de cómo podría desarrollarse un email marketing diferenciado por personas. Así, se concluye que lo expuesto en este trabajo permite dar por cumplido el principal objetivo del estudio.

9.2 Limitantes

La primera limitante es que la principal base de datos usada en este trabajo, la Base de envíos, refleja que el registro de los correos que ha enviado el banco, comenzó a implementarse tiempo después de que la empresa realizara email marketing. Más aún, esta información se regularizó para mediados del año 2020. Por ende, al tratarse de una base de datos algo nueva, perjudica la validez que pueden tener los resultados de este trabajo.

Otra limitante fue el prolongado periodo de inducción implementado por la empresa, el cual duró cerca de un mes y medio, retrasando el comienzo del trabajo. Ante esto, la memoria tuvo una

constante demora respecto a lo planificado, y la opción de llevar a cabo experimentación fue perdiendo terreno, imposibilitando la obtención de resultados más claros.

La última limitante son los diferentes factores que pueden alterar la lectura de correos por parte de las personas, y que son independientes del accionar de la organización, como por ejemplo problemas en los correos electrónicos o algún evento desafortunado de fuerza mayor por el que los individuos ignoran sus mails.

9.3 Trabajos futuros

Un claro trabajo futuro es la realización de los experimentos diseñados en una de las secciones previas, lo cual permitiría establecer una única política de toques a utilizar para contactar a cada persona.

Se plantea extender este trabajo para aquellos horarios que el banco no suele utilizar, para lo cual se hace necesario implementar con mayor reiteración el email marketing de la organización los días sábado y domingo, y dentro del rango 19:00 hrs. a 07:00 hrs. Además, esta memoria puede ser complementada con estudios sobre límites de correos a enviar por persona diaria o semanalmente, y sobre el tiempo que debe transcurrir entre envíos de mails.

También se propone seguir investigando formas para mejorar el email marketing de la empresa, el cual puede convertirse en el canal que mayor beneficios entregue a la organización considerando lo práctico y eficiente que puede llegar a ser.

Además, se sugiere determinar el medio por el que cada cliente prefiere ser contactado. De esta forma, el banco podría ajustar su *emailing* para aquellas personas que se inclinen por este canal de comunicación.

Por último, este estudio puede servir como una guía para realizar trabajos análogos en otros canales de comunicación y para diferentes grupos de personas o clientes.

10. Bibliografía

1. Agrupación jerárquica – *Hierarchical clustering*. Wikioes [en línea]: <https://upwikies.top/wiki/Hierarchical_Clustering> [Última consulta: 10 de febrero del 2022].

2. Cárdenas, Julián. (2014). "Regresión logística binaria". Networkianos.

3. Coeficiente de Gini. Wikipedia [en línea]: <https://es.wikipedia.org/wiki/Coeficiente_de_Gini> [Última consulta: 10 de febrero del 2022].

4. CRISP-DM: La metodología para poner orden a los proyectos. Sngular [en línea]: <https://www.sngular.com/es/data-science-crisp-dm-metodologia/> [Última consulta: 11 de febrero del 2022].

5. Dávila, Jorge. "Spam y su Regulación en Chile: ¿Cómo obtener la Protección Jurídica de la intimidad de las personas, sin afectar el desarrollo legítimo de una actividad económica?. Tesis de Magíster en Derecho Informático y de las Telecomunicaciones. Facultad de Derecho, Universidad de Chile. Santiago de Chile, 2011.

6. Diferencia de proporciones. Universidad de Barcelona [en línea]: <http://www.ub.edu/aplica_infor/spss/cap4-5.htm> [Última consulta: 23 de febrero del 2022].

7. Diferencia entre estandarización y normalización. ICHI.PRO [en línea]: <https://ichi.pro/es/diferencia-entre-estandarizacion-y-normalizacion> [Última consulta: 14 de febrero del 2022].

8. Döring, Matthias. (2018). "*Interpreting Generalized Linear Models*". Data Science Blog.

9. Duò, Matteo. (2021). "Más de 20 Estadísticas Imprescindibles sobre el Marketing por Email". Kinsta.

10. Dupouy, Carlos. "Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile". Tesis para optar al grado de Magister en Finanzas. Postgrado de Economía y Negocios, Universidad de Chile. Santiago de Chile, 2014.

11. Emma. (2018). "*Where we are & where we're going*". *Industry Report.*

12. Galán, Víctor. "Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario". Proyecto Fin de Carrera. Ingeniería en Informática, Escuela Politécnica Superior, Universidad Carlos III de Madrid. España, 2015.

13. González, Ligdi. (2020). "Algoritmo Agrupamiento Jerárquico - Teoría". AprendeIA.

14. Gutiérrez, Miguel. “Estudio de la saturación en email marketing para un negocio de retail”. Memoria para optar al título de Ingeniero Civil Industrial. Departamento de Ingeniería Civil Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago de Chile, 2019.

15. K vecinos más próximos. Wikipedia [en línea]:
<https://es.wikipedia.org/wiki/K_vecinos_más_próximos>
[Última consulta: 10 de febrero del 2021].

16. Ley Fácil - Spam. Biblioteca del Congreso Nacional de Chile [en línea]:
<https://www.bcn.cl/leyfacil/recurso/spam>
[Última consulta: 12 de febrero del 2022].

17. Ley 19.496. Biblioteca del Congreso Nacional de Chile [en línea]:
<https://www.bcn.cl/leychile/navegar?idNorma=61438>
[Última consulta: 12 de febrero del 2022].

18. Lizares, Mónica. “Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico”. Tesina para optar al Título Profesional de Licenciada en Estadística. Escuela Nacional de Estadística, Facultad de Ciencias Matemáticas. Lima, Perú, 2017.

19. Los 6 mejores bancos en Chile. PaySpace Magazine [en línea]:
<https://payspacemagazine.com/banks/los-6-mejores-bancos-en-chile/>
[Última consulta: 9 de febrero del 2022].

20. Martínez de Lejarza, Ignacio. “Árboles de clasificación y regresión”. Material del Máster en Ciencias Actuariales y Financieras, Universidad de Valencia. España.

21. Melillanca, Eric. (2018). “Noción de R-cuadrado o Coeficiente de Determinación”. Welcome to the Jungle.

22. Memoria Anual 2020. Sitio web banco [en línea]:
<https://www.bci.cl/investor-relations/memoria-anual/files/memoria-anual-2020>
[Última consulta: 9 de febrero del 2022].

23. Modelos saturados, desviación y derivación de la suma de cuadrados. ICHI.PRO [en línea]:
<https://ichi.pro/es/modelos-saturados-desviacion-y-derivacion-de-la-suma-de-cuadrados>
[Última consulta: 10 de febrero del 2022].

24. Montero, Roberto. “Modelos de regresión lineal múltiple”. Documento de Trabajo en Economía Aplicada. Departamento de Economía Aplicada, Universidad de Granada. España, 2016.

25. Murillo, Andrés. “Incorporación de técnicas de persuasión en email marketing”. Memoria para optar al título de Ingeniero Civil Industrial. Departamento de Ingeniería Civil Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago de Chile, 2016.

26. No Molestar. Sernac [en línea]:
<https://www.sernac.cl/portal/618/w3-propertyvalue-62998.html>
[Última consulta: 12 de febrero del 2022].

27. Otzen, Tamara & Manterola, Carlos. (2017). "Técnicas de Muestreo sobre una Población a Estudio". International Journal of Morphology. 35(1), 227-232.

28. ¿Qué es el email marketing y para qué sirve?. Foxize [en línea]:
<https://www.foxize.com/blog/que-es-el-email-marketing-y-para-que-sirve/>
[Última consulta: 9 de febrero del 2022].

29. Quintela del Río, Alejandro. (2019). "Estadística Básica Edulcorada". Bookdown.

30. Quinto, Carla. (2022). "Mejores bancos de Chile 2022". Rankia.

31. Reul, Mariana. (2021). "¿Qué es el email marketing y cómo puede ayudar a aumentar tus conversiones?". Sendinblue.

32. Rojas, Andrea. "Estudio experimental de automatización de email marketing en un retail online". Memoria para optar al título de Ingeniera Civil Industrial. Departamento de Ingeniería Civil Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago de Chile, 2014.

33. Rus, Enrique. (2021). "Muestreo aleatorio". Economipedia.

34. *Safest Banks in Latin America 2020*. Global Finance [en línea]:
<https://d2tyltutevw8th.cloudfront.net/media/document/press-release-safest-banks-by-region-2020-latin-1602791952.pdf>
[Última consulta: 9 de febrero del 2022].

35. Segmentación utilizando K-means en Python. Machine Learning Para Todos [en línea]:
<https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/>
[Última consulta: 10 de febrero de 2022].

36. Test para comparación de medias. Universidad de La Plata [en línea]:
<http://www.mate.unlp.edu.ar/practicas/55_8_30052010143027.pdf>
[Última consulta: 23 de febrero de 2022].

37. Vandeput, Nicolas. (2019). "*Forecast KPIs: RMSE, MAE, MAPE & Bias*". Towards Data Science.

38. Zamora, Andrea. (2016). "Nuevo proyecto de ley para evitar el spam de las empresas". IDA Blog.

39. *2021 Content Marketing B2C*. Content Marketing Institute [en línea]:
<https://contentmarketinginstitute.com/wp-content/uploads/2021/01/b2c-research-report-2021.pdf>
[Última consulta: 9 de febrero del 2022].

11.Anexos

Anexo A: Cantidad de envíos y tasa de apertura por tipo de correo

Tabla 25: Apertura de correos según tipo.
Fuente: Elaboración propia con datos de la Base de correos.

Ámbito	Tipo	Nº correos enviados [Millones]	Porción de correos abiertos
Fidelizar	Actualizar datos	0,09	20,06%
	Habilitación	1,30	25,48%
	Inversión	1,52	23,25%
	Onboarding	0,07	32,07%
	PAT	0,48	20,39%
	Tarjeta	39,18	23,55%
Vender	Aumento cupo	0,66	30,55%
	Avance	2,58	34,64%
	Consumo	2,66	28,04%
	Cuotización	2,02	26,42%
	Hipotecario	1,09	29,80%
	Inversión	0,90	27,30%
	Seguro	3,57	22,29%
	Tarjeta	0,22	31,43%
Informar	Canales digitales	10,08	22,92%
	News	2,42	26,26%
	Otros	6,74	25,25%
Cobrar	Riesgo	0,17	35,18%
Atraer	Planes	0,03	28,04%
Pyme	Pyme	0,03	28,04%

Anexo B: Apertura y cantidad de correos enviados por hora

Tabla 26: Apertura de correos según hora de envío.
Fuente: Elaboración propia con datos de la Base de correos.

Hora	Nº correos enviados [Miles]	Porción de correos abiertos
00:00:00-00:59:59	0,06	16,39%
01:00:00-01:59:59	0,04	74,81%
02:00:00-02:59:59	0,52	23,08%
03:00:00-03:59:59	1,28	29,69%
04:00:00-04:59:59	294,25	25,61%
05:00:00-05:59:59	5,93	24,96%
06:00:00-06:59:59	1.020,90	25,40%
07:00:00-07:59:59	2.059,11	20,03%
08:00:00-08:59:59	4.047,26	24,36%
09:00:00-09:59:59	5.206,81	24,64%
10:00:00-10:59:59	5.101,95	23,01%
11:00:00-11:59:59	8.234,32	23,26%
12:00:00-12:59:59	13.109,34	25,33%
13:00:00-13:59:59	11.885,12	24,88%
14:00:00-14:59:59	12.818,30	25,02%
15:00:00-15:59:59	8.670,59	24,66%
16:00:00-16:59:59	2.687,13	24,73%
17:00:00-17:59:59	655,32	26,52%
18:00:00-18:59:59	0,13	61,54%
19:00:00-19:59:59	0,06	16,81%
20:00:00-20:59:59	0,06	49,67%
21:00:00-21:59:59	0,09	55,51%
22:00:00-22:59:59	0,09	55,82%
23:00:00-23:59:59	0,06	50,17%

Anexo C: Composición de periodos por división

Tabla 27: Meses incluidos por periodo en cada división.
Fuente: Elaboración propia.

		2020						2021								
División	**Periodo**	**Jul.**	**Ago.**	**Sep.**	**Oct.**	**Nov.**	**Dic.**	**Ene.**	**Mar.**	**Abr.**	**May.**	**Jun.**	**Jul.**	**Ago.**	**Sep.**	**Oct.**
1	Primero	x	x	x	x											
	Segundo					x	x	x	x							
	Tercero									x	x	x	x			
2	Primero		x	x	x	x										
	Segundo						x	x	x	x						
	Tercero										x	x	x	x		
3	Primero			x	x	x	x									
	Segundo							x	x	x	x					
	Tercero											x	x	x	x	
4	Primero				x	x	x	x								
	Segundo								x	x	x	x				
	Tercero												x	x	x	x

Anexo D: Matrices inicio/caída e inicio/recuperación

Tabla 28: Matriz inicio/caída.
Fuente: Elaboración propia.

INICIO/CAÍDA	[0%;10%[	[10%;20%[	[20%;30%[	[30%;40%[	[40%;50%[	[50%;60%[	[60%;70%[	[70%;80%[	[80%;90%[	[90%;100%]	TOTAL
[0%; 10%[	3.053	2.013	2.051	1.519	564	0	0	0	0	0	9.201
[10%; 20%[	1.760	2.292	3.312	4.311	5.056	4.814	2.944	457	0	0	24.946
[20%; 30%[	589	922	1.479	2.152	2.838	3.381	3.919	3.556	297	0	19.133
[30%; 40%[	244	399	765	1.169	1.736	2.275	2.738	3.081	1.652	0	14.059
[40%; 50%[	158	247	400	715	1.036	1.586	1.946	2.357	2.124	0	10.569
[50%; 60%[	107	166	249	411	731	1.098	1.437	1.904	2.108	183	8.394
[60%; 70%[	56	83	148	280	446	715	1.058	1.501	1.864	455	6.606
[70%; 80%[	13	60	101	181	305	464	755	1.140	1.813	717	5.548
[80%; 90%[	1	24	77	122	204	266	501	960	1.817	1.179	5.150
[90%; 100%]	0	3	26	113	123	229	358	690	1.997	2.910	6.448
TOTAL	5.980	6.209	8.607	10.973	13.038	14.828	15.656	15.647	13.671	5.445	

Tabla 29: Matriz inicio/recuperación.
Fuente: Elaboración propia.

INICIO/RECUPERA	[0%;10%[	[10%;20%[	[20%;30%[	[30%;40%[	[40%;50%[	[50%;60%[	[60%;70%[	[70%;80%[	[80%;90%[	[90%;100%]	>100%	TOTAL
[0%; 10%[	0	0	0	0	0	6	45	69	190	197	1.511	2.018
[10%; 20%[	0	0	3	25	101	230	365	541	733	844	4.160	7.002
[20%; 30%[	0	1	22	68	149	270	393	513	685	858	3.103	6.062
[30%; 40%[	0	3	20	68	117	209	309	439	530	683	2.242	4.620
[40%; 50%[	0	7	21	40	70	148	230	348	506	577	1.640	3.588
[50%; 60%[	0	4	20	32	48	111	163	257	372	522	1.258	2.787
[60%; 70%[	0	2	7	14	30	62	119	192	327	475	1.033	2.261
[70%; 80%[	0	0	4	14	15	38	82	133	279	502	784	1.851
[80%; 90%[	0	0	0	2	12	22	39	93	205	531	697	1602
[90%; 100%]	0	0	0	2	3	9	18	42	152	906	621	1.754
TOTAL	0	17	98	266	546	1.106	1.763	2.628	3.979	6.094	17.048	

Anexo E: Coeficientes de silueta para métodos de segmentación

Tabla 30: Siluetas para Agrupamiento jerárquico aglomerativo.
Fuente: Elaboración propia.

	Muestras										
Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Promedio**
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,60	0,59	0,59	0,60	0,60	0,60	0,61	0,59	0,60	0,60	0,60
3	0,31	0,35	0,27	0,38	0,38	0,33	0,37	0,35	0,37	0,37	0,35
4	0,32	0,36	0,23	0,40	0,39	0,34	0,24	0,25	0,22	0,38	0,31
5	0,27	0,21	0,24	0,23	0,22	0,23	0,26	0,26	0,23	0,23	0,24
6	0,24	0,20	0,22	0,17	0,25	0,20	0,23	0,16	0,18	0,22	0,21
7	0,23	0,17	0,21	0,19	0,21	0,19	0,21	0,17	0,18	0,18	0,19
8	0,23	0,15	0,20	0,17	0,18	0,20	0,17	0,15	0,19	0,19	0,18
9	0,20	0,15	0,21	0,16	0,18	0,17	0,16	0,16	0,18	0,19	0,18
10	0,19	0,13	0,21	0,15	0,19	0,18	0,17	0,16	0,18	0,17	0,17

Tabla 31: Siluetas para *K*-medias.
Fuente: Elaboración propia.

	Muestras										
Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Promedio**
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,30	0,29	0,29	0,29	0,38	0,29	0,61	0,29	0,31	0,29	0,33
3	0,37	0,36	0,36	0,37	0,37	0,36	0,39	0,36	0,38	0,36	0,37
4	0,24	0,25	0,25	0,24	0,30	0,24	0,18	0,25	0,27	0,24	0,25
5	0,22	0,22	0,23	0,22	0,27	0,21	0,22	0,21	0,23	0,24	0,23
6	0,27	0,22	0,24	0,20	0,24	0,24	0,23	0,21	0,21	0,24	0,23
7	0,23	0,18	0,22	0,20	0,22	0,20	0,24	0,21	0,22	0,22	0,22
8	0,23	0,21	0,25	0,19	0,20	0,18	0,21	0,19	0,19	0,18	0,20
9	0,23	0,19	0,24	0,21	0,23	0,22	0,18	0,20	0,20	0,18	0,21
10	0,19	0,19	0,24	0,18	0,19	0,21	0,20	0,19	0,22	0,19	0,20

Anexo F: Cambio de parámetros para partición

Variable	Partición
Abierto	70%
Abierto	80%
Día de envío	70%
Día de envío	80%
Hora de envío	70%
Hora de envío	80%

Figura 22: Secuencia de cambios en parámetros para partición de datos en Base de correos.
Fuente: Elaboración propia.

Anexo G: Horarios de envío

Tabla 32: Horarios de envío según clase de correo y segmento de clientes.
Fuente: Elaboración propia.

Fidelizar Actualizar datos	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Lunes	12:00:00 - 12:59:59	Lunes	11:00:00 - 11:59:59
BL	Jueves	11:00:00 - 11:59:59	Lunes	11:00:00 - 11:59:59
SNL	Lunes	12:00:00 - 12:59:59	-	-
AL	Lunes	14:00:00 - 14:59:59	Lunes	13:00:00 - 13:59:59
SL	Lunes	15:00:00 - 15:59:59	-	-
ML	Lunes	14:00:00 - 14:59:59	-	-

Fidelizar Habilitación	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Domingo	6:00:00 - 6:59:59	Domingo	3:00:00 - 3:59:59
BL	Sábado	6:00:00 - 6:59:59	Sábado	3:00:00 - 3:59:59
SNL	Lunes	7:00:00 - 7:59:59	-	-
AL	Martes	3:00:00 - 3:59:59	Lunes	3:00:00 - 3:59:59
SL	Sábado	6:00:00 - 6:59:59	-	-
ML	Martes	3:00:00 - 3:59:59	-	-

Fidelizar Inversión	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	21:00:00 - 21:59:59	-	-
BL	Lunes	7:00:00 - 7:59:59	-	-
SNL	Jueves	9:00:00 - 9:59:59	-	-
AL	Lunes	15:00:00 - 15:59:59	Lunes	16:00:00 - 16:59:59
SL	Viernes	8:00:00 - 8:59:59	-	-
ML	Lunes	11:00:00 - 11:59:59	-	-

Fidelizar Onboarding	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Jueves	16:00:00 - 16:59:59	-	-
BL	Miércoles	15:00:00 - 15:59:59	-	-
SNL	Viernes	10:00:00 - 10:59:59	-	-
AL	Jueves	16:00:00 - 16:59:59	-	-
SL	Jueves	16:00:00 - 16:59:59	-	-
ML	Jueves	16:00:00 - 16:59:59	-	-

Fidelizar PAT	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Lunes	12:00:00 - 12:59:59	-	-
BL	Jueves	14:00:00 - 14:59:59	Lunes	14:00:00 - 14:59:59
SNL	Martes	14:00:00 - 14:59:59	-	-
AL	Martes	13:00:00 - 13:59:59	-	-
SL	Martes	13:00:00 - 13:59:59	Lunes	13:00:00 - 13:59:59
ML	Jueves	14:00:00 - 14:59:59	Jueves	13:00:00 - 13:59:59

Fidelizar Tarjeta	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Lunes	17:00:00 - 17:59:59	-	-
BL	Jueves	17:00:00 - 17:59:59	-	-
SNL	Lunes	12:00:00 - 12:59:59	-	-
AL	Martes	17:00:00 - 17:59:59	-	-
SL	Viernes	6:00:00 - 6:59:59	-	-
ML	Jueves	17:00:00 - 17:59:59	-	-

Vender Aumento cupo	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Viernes	17:00:00 - 17:59:59	-	-
BL	Viernes	16:00:00 - 16:59:59	-	-
SNL	Lunes	9:00:00 - 9:59:59	-	-
AL	Viernes	17:00:00 - 17:59:59	-	-
SL	Viernes	12:00:00 - 12:59:59	-	-
ML	Viernes	17:00:00 - 17:59:59	-	-

Vender Avance	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Viernes	14:00:00 - 14:59:59	-	-
BL	Miércoles	14:00:00 - 14:59:59	-	-
SNL	Miércoles	8:00:00 - 8:59:59	-	-
AL	Viernes	15:00:00 - 15:59:59	-	-
SL	Miércoles	7:00:00 - 7:59:59	-	-
ML	Viernes	15:00:00 - 15:59:59	-	-

Vender Consumo	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	10:00:00 - 10:59:59	-	-
BL	Martes	12:00:00 - 12:59:59	-	-
SNL	Martes	7:00:00 - 7:59:59	-	-
AL	Jueves	14:00:00 - 14:59:59	-	-
SL	Martes	15:00:00 - 15:59:59	-	-
ML	Miércoles	16:00:00 - 16:59:59	Jueves	16:00:00 - 16:59:59

Vender Cuotización	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	14:00:00 - 14:59:59	-	-
BL	Jueves	14:00:00 - 14:59:59	-	-
SNL	Viernes	4:00:00 - 4:59:59	-	-
AL	Martes	9:00:00 - 9:59:59	-	-
SL	Lunes	8:00:00 - 8:59:59	-	-
ML	Martes	16:00:00 - 16:59:59	-	-

Vender Hipotecario	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	6:00:00 - 6:59:59	-	-
BL	Martes	10:00:00 - 10:59:59	-	-
SNL	Jueves	15:00:00 - 15:59:59	-	-
AL	Martes	8:00:00 - 8:59:59	-	-
SL	Jueves	14:00:00 - 14:59:59	-	-
ML	Martes	14:00:00 - 14:59:59	-	-

Vender Inversión	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	6:00:00 - 6:59:59	Lunes	6:00:00 - 6:59:59
BL	Martes	6:00:00 - 6:59:59	Viernes	6:00:00 - 6:59:59
SNL	Miércoles	15:00:00 - 15:59:59	-	-
AL	Martes	6:00:00 - 6:59:59	-	-
SL	Viernes	8:00:00 - 8:59:59	-	-
ML	Lunes	11:00:00 - 11:59:59	Lunes	6:00:00 - 6:59:59

Vender Seguro	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	15:00:00 - 15:59:59	-	-
BL	Lunes	7:00:00 - 7:59:59	-	-
SNL	Miércoles	9:00:00 - 9:59:59	-	-
AL	Martes	18:00:00 - 18:59:59	-	-
SL	Miércoles	11:00:00 - 11:59:59	-	-
ML	Martes	7:00:00 - 7:59:59	-	-

Vender Tarjeta	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Jueves	12:00:00 - 12:59:59	-	-
BL	Jueves	12:00:00 - 12:59:59	Jueves	10:00:00 - 10:59:59
SNL	Miércoles	8:00:00 - 8:59:59	-	-
AL	Viernes	10:00:00 - 10:59:59	-	-
SL	Jueves	12:00:00 - 12:59:59	Lunes	12:00:00 - 12:59:59
ML	Viernes	10:00:00 - 10:59:59	-	-

Informar Canales digitales	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Domingo	13:00:00 - 13:59:59	Domingo	12:00:00 - 12:59:59
BL	Domingo	13:00:00 - 13:59:59	-	-
SNL	Miércoles	15:00:00 - 15:59:59	-	-
AL	Domingo	14:00:00 - 14:59:59	Domingo	15:00:00 - 15:59:59
SL	Domingo	14:00:00 - 14:59:59	-	-
ML	Domingo	14:00:00 - 14:59:59	Domingo	15:00:00 - 15:59:59

Informar News	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Viernes	9:00:00 - 9:59:59	-	-
BL	Viernes	9:00:00 - 9:59:59	-	-
SNL	Viernes	8:00:00 - 8:59:59	-	-
AL	Miércoles	16:00:00 - 16:59:59	-	-
SL	Miércoles	9:00:00 - 9:59:59	-	-
ML	Martes	15:00:00 - 15:59:59	-	-

Informar Otros	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Viernes	7:00:00 - 7:59:59	Lunes	7:00:00 - 7:59:59
BL	Viernes	7:00:00 - 7:59:59	-	-
SNL	Lunes	8:00:00 - 8:59:59	-	-
AL	Lunes	8:00:00 - 8:59:59	Domingo	7:00:00 - 7:59:59
SL	Martes	6:00:00 - 6:59:59	Lunes	6:00:00 - 6:59:59
ML	Viernes	7:00:00 - 7:59:59	Domingo	7:00:00 - 7:59:59

Cobrar Riesgo	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Lunes	8:00:00 - 8:59:59	-	-
BL	Lunes	8:00:00 - 8:59:59	-	-
SNL	Lunes	8:00:00 - 8:59:59	-	-
AL	Lunes	13:00:00 - 13:59:59	-	-
SL	Viernes	7:00:00 - 7:59:59	-	-
ML	Lunes	13:00:00 - 13:59:59	-	-

Atraer Planes	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Martes	1:00:00 - 1:59:59	-	-
BL	Martes	1:00:00 - 1:59:59	-	-
SNL	Martes	10:00:00 - 10:59:59	-	-
AL	Martes	12:00:00 - 12:59:59	-	-
SL	Domingo	11:00:00 - 11:59:59	-	-
ML	Miércoles	10:00:00 - 10:59:59	-	-

Pyme Pyme	Recomendado		Propuesto	
Segmento	**Día**	**Hora**	**Día**	**Hora**
General	Lunes	13:00:00 - 13:59:59	-	-
BL	Martes	15:00:00 - 15:59:59	-	-
SNL	Martes	13:00:00 - 13:59:59	-	-
AL	Jueves	10:00:00 - 10:59:59	-	-
SL	Lunes	8:00:00 - 8:59:59	-	-
ML	Jueves	10:00:00 - 10:59:59	-	-