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

DISEÑO EXPERIMENTAL PARA DETERMINAR DONANTES INFLUENCIABLES POR CAMPAÑAS COMUNICACIONALES APLICADO A LA CARTERA DE SOCIOS DE FUNDACIÓN LAS ROSAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DIEGO JAVIER ALFARO BÁEZ

PROFESOR GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
PABLO GONZÁLEZ MALEBRÁN

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Diego Javier Alfaro Báez
FECHA: 20/01/2020
PROFESOR GUÍA: Alejandra Puente Chandía

Diseño experimental para determinar donantes influenciables por campañas comunicacionales aplicado a la cartera de socios de Fundación las Rosas

Fundación las Rosas es una institución privada, sin fines de lucro, católica, que acoge a los adultos mayores más vulnerables del país. En sus 29 hogares residen 2.200 adultos mayores. El financiamiento de toda la operación se da a través de distintos canales, siendo el Programa Amigos el más importante al representar cerca del 40% de los ingresos.

El Programa Amigos es el medio por el cual las personas pueden realizar donanciones mensuales con cargo a cuentas bancarias. Actualmente hay más de 100.000 afiliados en este programa y genera ingresos por más de 6 mil millones de pesos anuales.

La comunicación con los afiliados se da principalmente por correo electrónico, por este medio se les invita a participar en distintas campañas que buscan aumentar los ingresos del programa. La realización de estas campañas no incluye un posterior evaluación de la misma, no existe un grupo de control, por lo que no se pueden obtener conclusiones fiables. Por otra parte, no se considera que puede existir un efecto negativo en algunos segmentos de la cartera, lo que puede producir la fuga de estos.

Por lo anterior, se propone mejorar la eficiencia de las campañas mediante la implementación de un modelo uplift el cual se alimentará de un experimento de campo. El diseño del experimento se realizará en base a un análisis descriptivo y a modelos predictivos, además considerando la utilidad que puedan tener para la institución.

Dentro de los principales hallazgos está la relación cuadrática que existe entre el monto de aporte y la propensión a aceptar un aumento, así, para hay una relación directa hasta un punto de inflexión del aporte en el que empiezan a ser menos propensos a aumentar las personas con mayores valores de aporte.

Se aprecia además la diferencia entre querer predecir un valor de aumento versus querer predecir si la persona aumentará o no. Esta diferencia se vio al comparar modelos de regresión logística y numérica y por las cantidades requeridas para los experimentos.

TABLA DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

1. Antecedentes Generales

1.1. Caracterización de la Institución

Fundación las Rosas (FLR) es una institución privada, católica, sin fines de lucro, que se dedica al cuidado de los adultos mayores con mayor nivel de vulnerabilidad socio-económica en el país. Actualmente cuenta con 28 hogares ubicados en 8 regiones del país, en ellos residen más de 2.200 adultos mayores que están al cuidado de más de 1.500 personas, entre cuidadores y profesionales de la salud.[1, 2, 3 y 6]

Como institución, Fundación las Rosas declara la siguiente misión:

> *"Acoger, alimentar, acompañar en la salud y en el encuentro con el Señor a personas mayores pobres y desvalidas, manteniéndolas integradas a la familia y a la sociedad de forma digna y activa."*

Con esto se quiere asegurar el buen cuidado de sus residentes, una buena alimentación, buena salud, y todas las atenciones que ancianos, mayoritariamente enfermos, puedan necesitar. Se incluye, también, el afán de mantenerlos activos, estimularlos con distintas actividades y generar instancias adecuadas para estimularlos cognitivamente.

Complementariamente, su visión es:

> *"Ser, como institución de la Iglesia Católica, una fuente de inspiración y testimonio de amor y servicio a las personas mayores."*

Lo anterior se ve reflejado en la importancia que se le da, en el cuidado del adulto mayor, al acompañamiento espiritual. Los ancianos que llegan a la Fundación saben que pasarán ahí sus últimos años de vida, por lo que en este cuidado se espera que puedan ser felices, que se sientan acompañados y así, puedan morir en paz y en compañía de Dios. Es por esto, además, que en la administración de los hogares se busca que trabajen monjas de alguna congregación religiosa que puedan ser las guías en este acompañamiento.

1.2. La Estructura Organizacional

Como se puede ver en la Ilustración 1, organizacionalmente Fundación las Rosas deja en la cabeza al Vicepresidente Ejecutivo y Capellán, le sigue el Directorio y la Capellanía, luego el Gerente General, que tiene a cargo 4 direcciones organizadas funcionalmente y 4 áreas más pequeñas que las direcciones. Las 4 direcciones son la Dirección de Desarrollo, Dirección de Gestión de Hogares, Dirección de Personas y Dirección de Administración y Proyectos. Las 4 áreas son Auditoría, Legal, Gestión Cultural Institucional y Formación y Capacitación.

Ilustración 1 - Organigrama Fundación las Rosas

Fuente: Nosotros – Fundación las Rosas [3]

Las cuatro direcciones se crearon por los distintos objetivos en las funciones que realizan, así, la Dirección de Desarrollo está encargada, principalmente, de generar los ingresos que no se reciben desde el Estado para el funcionamiento de los hogares. La Dirección de Gestión de Hogares está a cargo de la administración y mantención de los hogares. La Dirección de Personas está a cargo del manejo de recursos humanos dentro de la Fundación. Finalmente, la Dirección de Administración y Proyectos está a cargo de la contabilidad y el abastecimiento de todas las unidades de la Fundación.

Fundación las Rosas lleva 51 años de labor, en este tiempo se ha vivido un crecimiento acelerado en la cantidad de plazas disponibles para acoger adultos mayores. Este crecimiento no fue acompañado, en su momento, de la gestión y los cambios organizacionales necesarios. Esto provoca, en la actualidad, gran parte de las ineficiencias, debido principalmente por los sistemas informáticos que se desarrollaron, ya que son sistemas que no se comunican entre ellos, están poco automatizados y, en algunos casos, no tienen el desempeño óptimo ya que fueron diseñados para el manejo de bases de datos más pequeñas.

1.3. Marco Institucional

1.3.1. Mercado

De acuerdo a los resultados del Censo 2017, de las 17 millones de personas censadas, poco más 2 millones corresponden a personas de tercera edad. Por otra parte, considerando los resultados de la CASEN 2017, si de esas personas se considera que el 4,5% está en situación de pobreza por ingresos, se obtiene que, aproximadamente, 90 mil personas mayores está en situación de pobreza por ingresos. Estas personas reciben por parte del Estado pensiones solidarias, equivalentes a $107.304, monto que no alcanzan a cubrir todos los gastos que la vejez implica, por lo que, o representan un gasto para las familias, o no alcanzan a cubrir sus necesidades, o entran en algún ELEAM[1] social.[14 y 16]

Los ELEAM sociales son administrados principalmente por 4 entidades, cada una con coberturas diferentes; En primer lugar, está Fundación las Rosas con 2.200 plazas, luego la fundación San Vicente de Paul con 1.200, luego el Hogar de Cristo con 800 plazas para adultos mayores y CONAPRAM[2] con casi 500 plazas. Además, existen establecimientos independientes que suman cerca de 700 plazas. Todos estos establecimientos reciben apoyo y están bajo el alero del SENAMA[3]. Sumando todas estas plazas, se obtienen 5.400, destinadas para los 90 mil adultos mayores en situación de pobreza.[7]

Por lo anterior, es que existe una lista de espera de aproximadamente 1.300 personas, todas ellas cumplen con las condiciones de vulnerabilidad para ser acogidas por la Fundación. Al no existir disponibilidad de camas para poder acogerlos, es que se les da prioridad a las personas que están más enfermas, con mayor deterioro cognitivo y más dependientes, así, es que entran los adultos mayores más costosos, por lo que la generación de ingresos es un tema fundamental para la institución.

1.3.2. Marco Legislativo

Para que un ELEAM pueda operar, este debe estar aprobado por la Secretaría Regional Ministerial de Salud. Para obtener la aprobación, el ELEAM debe cumplir con requisitos mínimos expuestos por el Ministerio de Salud en el Decreto 14, el cual busca asegurar condiciones en las que se encontrarán los residentes, para evitar hacinamiento o malas condiciones de higiene y seguridad. Así, exige las cantidades máximas de residentes, dependiendo del tamaño del hogar, el personal necesario para el cuidado de los adultos mayores, dependiendo de su nivel de valencia, y hace exigencias en cuanto a infraestructura y funcionamiento de los hogares. La Secretaría Regional Ministerial de Salud es, además, la encargada de fiscalizar que se sigan cumpliendo estos requerimientos. [8]

[1] Establecimiento de Larga Estadía para Adultos Mayores
[2] Corporación Nacional de Protección de la Ancianidad
[3] Servicio Nacional del Adulto Mayor, Ministerio de Desarrollo Social

Por otra parte, al ser una institución sin fines de lucro, puede recibir donaciones las cuales pueden representar beneficios tributarios para los donantes, ya sean estas empresas o personas naturales, siempre que se cumplan las condiciones exigidas en la Ley del Impuesto a las Asignaciones y Donaciones. [9 y 10]

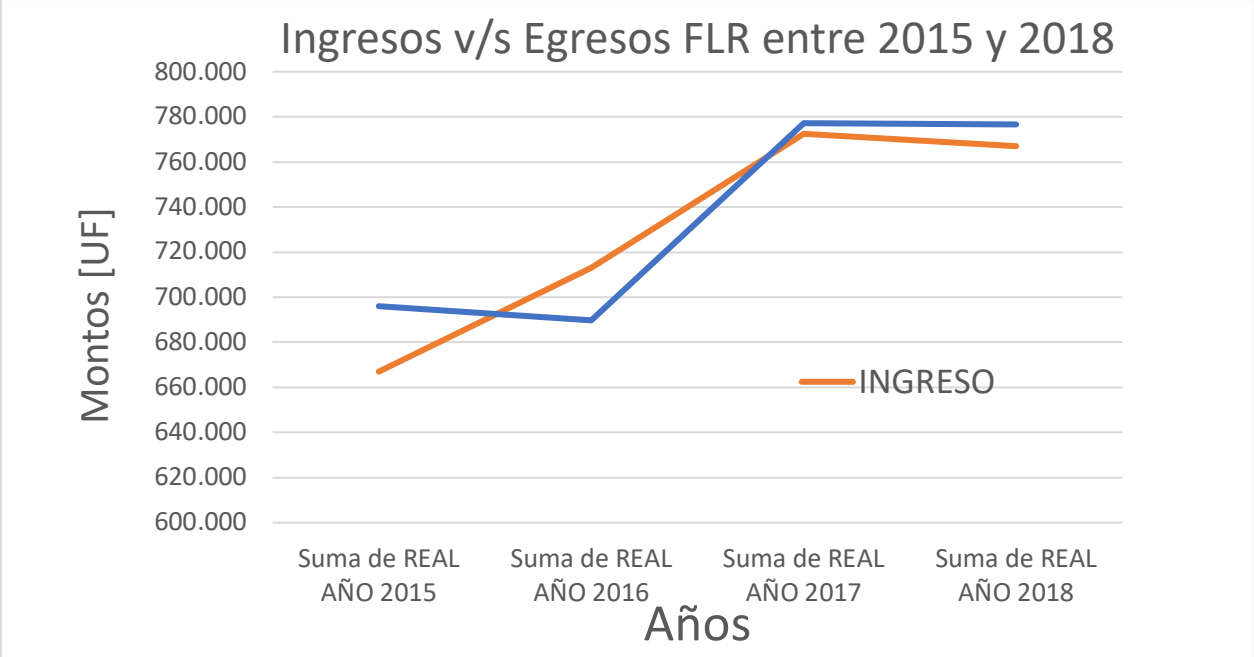
En Fundación las Rosas trabajan 1.800 personas, por lo tanto, debe regirse por el código del trabajo para todos los efectos relacionados con recursos humanos. Adicionalmente, en la institución existen 2 sindicatos, uno de ellos es exclusivo de las personas que trabajan en el Hogar 6 María Auxiliadora ubicado en Lampa. Para los demás colaboradores existe otro sindicato. [11]

1.4. Desempeño Organizacional

1.4.1. De los ingresos y costos

Como se mencionó anteriormente, durante los últimos años, los niveles de vulnerabilidad de los adultos mayores que residen en Fundación las Rosas han ido aumentando, están ingresando cada vez más enfermos, más dependientes y más abandonados. Esto provoca que los costos del cuidado hayan aumentado ya que se requiere de más horas hombre para la asistencia en las labores. Esto se ve al analizar el aumento en gastos en remuneraciones, desde el 2015 al 2018 aumentó en MM$3, esto es el 23,5% (ver Tabla 12 - Egresos por Ítem en 64). El gasto en remuneraciones representa del orden del 75% de los gastos de la Fundación, es por esto, que, contablemente, se hace relevante el estado de salud en el que se encuentran los residentes, ya que este afecta directamente en los egresos.

Gráfico 1 - Ingresos y egresos entre 2015 y 2018, Fuente: Elaboración propia

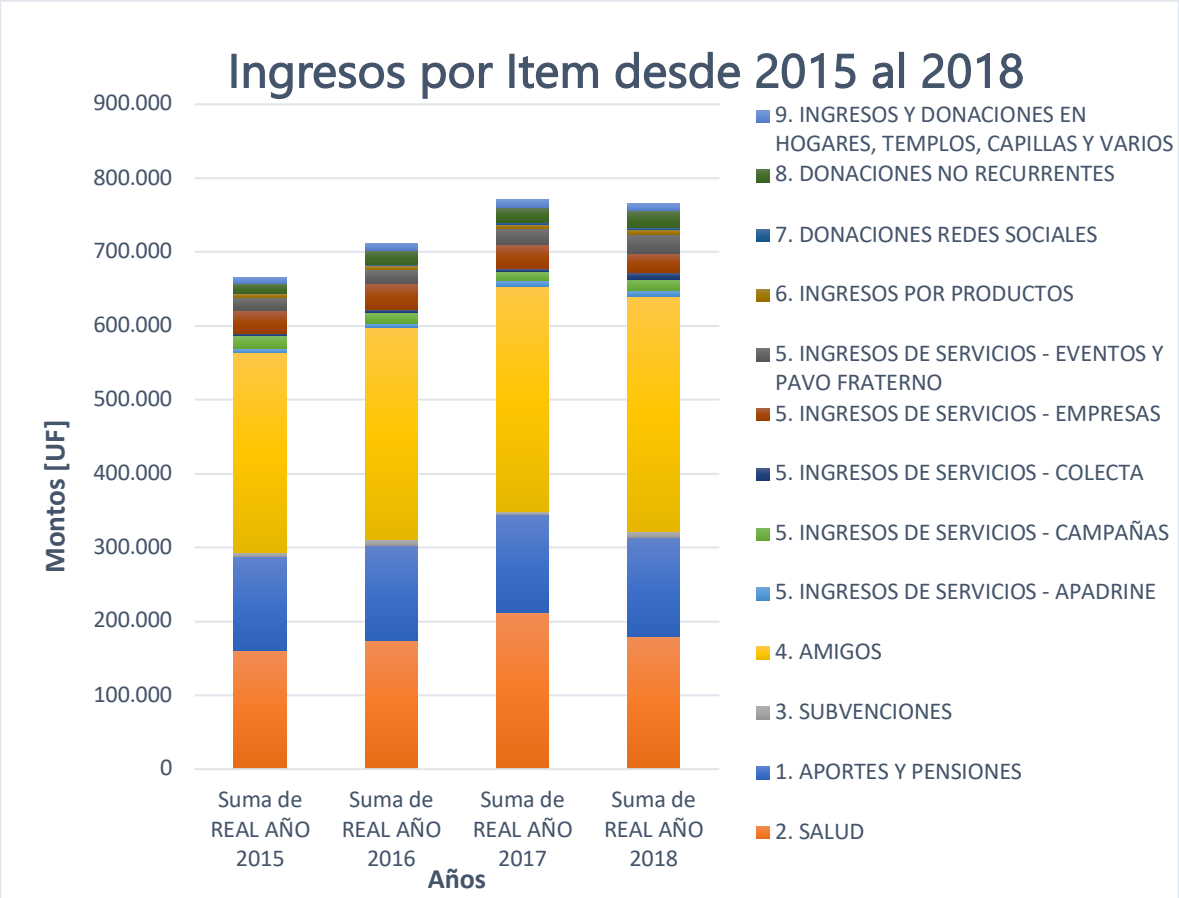


Fuente: Elaboración propia [12]

Así, frente a esta alza de los egresos, Fundación las Rosas se ha visto obligada a buscar fuentes de financiamiento, lo que ha llevado a poder cubrir estos gastos, pero siempre quedando al debe o con poco margen para seguir creciendo o invirtiendo (ver Gráfico 1).

Los ingresos vienen, principalmente, desde tres fuentes: pensiones de los residentes, cercano al 16%, ingresos por salud, entre SENAMA y FONASA[4], cercano al 25% y el canal de Amigos que aporta cerca del 40% de los ingresos de la Fundación (ver Gráfico 2). Las primeras dos fuentes de ingresos son externas, son subvenciones entregadas por el Estado en las que Fundación las Rosas no tiene mucho espacio para poder gestionar.

Gráfico 2 - Ingresos por Ítem



Fuente: Elaboración propia [12]

Por otra parte, el canal de Amigos incluye las donaciones que hacen personas naturales o pequeñas empresas, que están suscritas a la Fundación y que mensualmente se les realiza un cargo automático a sus cuentas bancarias. Este canal fue creado por Fundación las Rosas y actualmente es administrado

[4] Fondo Nacional de Salud

en la Dirección de Desarrollo en la que se trabaja constantemente por captar nuevos socios al programa de Amigos, fidelizar a los actuales, campañas de aumentos de aportes, entre otras gestiones.

Además, la Institución tiene otras líneas de ingresos, anualmente se realiza una Colecta Nacional (aporta cerca del 0,5%), una cena de Pavo Fraterno (~3%), un retail de productos llamado Bazar (~0,8%), el Apadrine, en el cual se conecta una empresa con un Hogar (1%), entre otras vías de ingresos que, porcentualmente, no generan mayor impacto.

1.4.2. De la administración de la cartera Amigos

La comunicación con los amigos inscritos en el programa está centralizada desde la Dirección de Desarrollo, no obstante, esta depende de 3 áreas: Comunicaciones y Marketing, Contact Center y Evaluación de Proyectos. Estas 3 áreas conversan y coordinan los e-mailing que se les envían (ver Ilustración 2), pero no existe una segmentación clara de los afiliados, por lo que siempre se trabaja bajo una misma línea comunicacional, seleccionando bases por atributos demográficos, en lugar de seleccionar según el tipo de amigo o vínculo que tiene la persona con FLR. Además de la planificación de mailings del año, siempre se pueden agregar nuevas informaciones si algún área lo considera pertinente, esto sin un protocolo claro y sin medir los posibles efectos negativos que tiene el envío de comunicaciones excesivo.

Ilustración 2 - Programación mailing amigos

	ENERO	FEBRERO	MARZO	ABRIL	MAYO	JUNIO
Amigos NUEVOS	Línea VÍNCULO NACIONAL	L. COMERCIAL NACIONAL	L. COMERCIAL NACIONAL	Línea VÍNCULO	Línea VÍNCULO	L. INFORMATIVA (News Nacional-regional)
	Mailing	Mailing	Mailing	Call center	Mailing invitación/regional	Mailing
	Buenos deseos para el año. Llamado RRSS	*SOAP (informativo)*	*SOAP (ventas)*	*Visibilizar ayuda*	*Ruta patrimonial Santuario. Visita Hogares*	*Aniversario, Colecta, Invierno, Nota regional, Post Aniversario*
KPI						
Meta						
Amigos VIGENTES	Línea VÍNCULO NACIONAL	Línea COMERCIAL	Línea COMERCIAL	Línea VÍNCULO	Línea VÍNCULO	L. INFORMATIVA (News Nacional-regional)
	Mailing	Mailing	Mailing	Mailing Nacional	Mailing invitación/regional	Aniversario, Colecta, Invierno, Nota regional, Post Aniversario
	Buenos deseos para el año. Llamado RRSS	*SOAP (informativo)*	*SOAP (ventas)*	*Transmitir nuestro foco – Visibilizar soledad*	*Ruta patrimonial Santuario. Visita Hogares*	*Aniversario – Colecta – Invierno – Nota regional – Post Aniversario*
KPI						
Meta						
Amigos INACTIVOS		L. COMERCIAL	L COMERCIAL	L. VÍNCULO		L. INFORMATIVA (News Nacional-regional)
		Mailing	Mailing	Mailing Nacional	Mailing invitación / regional	
		SOAP (informativo)	*SOAP (ventas)*	*Te extrañamos / soledad*	*Ruta patrimonial Santuario. Visita Hogares*	*Aniversario – Colecta – Invierno – Nota regional – Post Aniversario*
KPI						
Meta						

Fuente: Archivo interno

De la imagen anterior, se puede ver que existe un afán por establecer KPI's en los mailings que se envían, pero esto no se ha desarrollado hasta la fecha.

A modo de ejemplo, de la precaria segmentación, al momento de inscribir a una persona al Programa Amigos, dentro de todos los datos de contacto y pago, se le pregunta si desea recibir información de FLR, si desea ser voluntario en algún hogar de FLR y si acepta un reajuste de su aporte conforme al IPC. De estas 3 variables, sólo se rescata la del aumento de aporte, que se registra en una base independiente de donde se registra el resto de la información, y solo se utiliza esa información una vez al año para hacer los ajustes de montos. La información de si el amigo quiere recibir información o si quiere ser voluntario no queda registrada en ningún sistema ni se hace ninguna gestión al respecto. Al momento de analizar el vínculo que las personas tienen con FLR, la intuición dice que esas 3 variables pueden explicar mucho las diferencias que existen entre dos amigos del programa, pero esa información no se rescata y no se registra digitalmente.

1.4.3. De la imagen de marca

Al ser, el Canal Amigos, la principal fuente de ingresos de la Fundación, es importante analizar la percepción que tienen los donantes. Así, si se analiza el Estudio de Marcas Ciudadanas de Cadem, se puede ver que Fundación las Rosas está en el puesto 25 de 100, lo que la ubica en una muy buena posición y reconocida como nueva exitosa por los autores del estudio. No obstante, al separar por edad, es posible ver que, solo para las personas sobre 50 años, Fundación las Rosas es una marca relevante, mientras que los grupos más jóvenes incluyen, dentro de las instituciones sin fines de lucro, a Teletón, Desafío Levantemos Chile, TECHO y Hogar de Cristo. [13]

Además, en el estudio se detallan posiciones (hasta 200) según visibilidad, relevancia y aporte (ver Ilustración 17 - Dimensiones Evaluadas en 65), en estas categorías, Fundación las Rosas se encuentra posición 97 según visibilidad, 40 según relevancia y 7 según aporte. De aquí se puede deducir que, aunque Fundación las Rosas no está dentro de las marcas más visibles, la población sí la reconoce como un aporte a la sociedad, esto ya que el 61% de los encuestados reconoce que Fundación las Rosas realiza un aporte concreto a la sociedad.

De la categoría de relevancia (qué tan importante es la marca para la persona), se pueden deducir los principales competidores de Fundación las Rosas, considerando que debe competir por donantes. Estos son: Teletón, TECHO, Desafío Levantemos Chile y Hogar de Cristo.

Ilustración 3 - Campaña Colecta Nacional 2018

Fuente: Archivo interno

Ilustración 4 - Campaña Amigos 2018

Fuente: Archivo interno

Ilustración 5 - Campaña Navidad 2018

Fuente: Archivo interno

Haciendo un análisis desde otra mirada, al analizar las temáticas de las campañas comunicacionales se puede ver que están todas enmarcadas desde escenarios negativos, todas las campañas se centran en la pobreza, el abandono y la dependencia que los adultos mayores tienen. Esto refleja que no existe una mirada a largo plazo a la hora de crear campañas

comunicacionales, ya que estas campañas van aportando a la imagen de marca de FLR, al ser siempre negativas provocará que las personas asocien a FLR con sentimientos de tristeza, en lugar de sentimientos de esperanza y de descanso. Esto se provoca por el Efecto Halo, un sesgo cognitivo que generaliza una cualidad o un conjunto de atributos con la marca que está usando tales atributos, así al ver constantemente a FLR asociada a sentimientos negativos, se está provocando que, en el largo plazo, la marca FLR sea asociada a pensamientos negativos (ver Ilustración 3,Ilustración 4 yIlustración 5 - Campaña Navidad 2018). [17]

2. Justificación del Tema

2.1. Área que enmarca el proyecto

El proyecto está enmarcado en la Dirección de Desarrollo, la que está encargada, en primera instancia, de generar los ingresos que no se reciben desde el Estado. En esta Dirección hay cinco áreas (ver Ilustración 6), el área de Alianzas Corporativas es la encargada de generar vínculos con empresas, traer nuevas ideas desde ellas, proyectos que se puedan desarrollar, conseguir asesorías y/o apoyo de cualquier tipo. El área de Evaluación de Proyectos es la encargada de mejorar y hacer más eficiente los procesos de la Dirección, evaluar e implementar nuevos proyectos. El área de Operaciones está a cargo de lo que se realiza cotidianamente en la Dirección, esto es, venta de productos, administración del canal amigos y captación de nuevos amigos. Transversalmente está el área de Comunicaciones, que se encarga de la administración de las redes sociales, generar toda la grafía de la institución, mantener la imagen de marca y todo lo relacionado con comunicación de masas. Finalmente, el Contact Center que está a cargo de la comunicación individual con los socios de FLR y potenciales socios.

Ilustración 6 - Organigrama Dirección de Desarrollo

Fuente: Elaboración propia

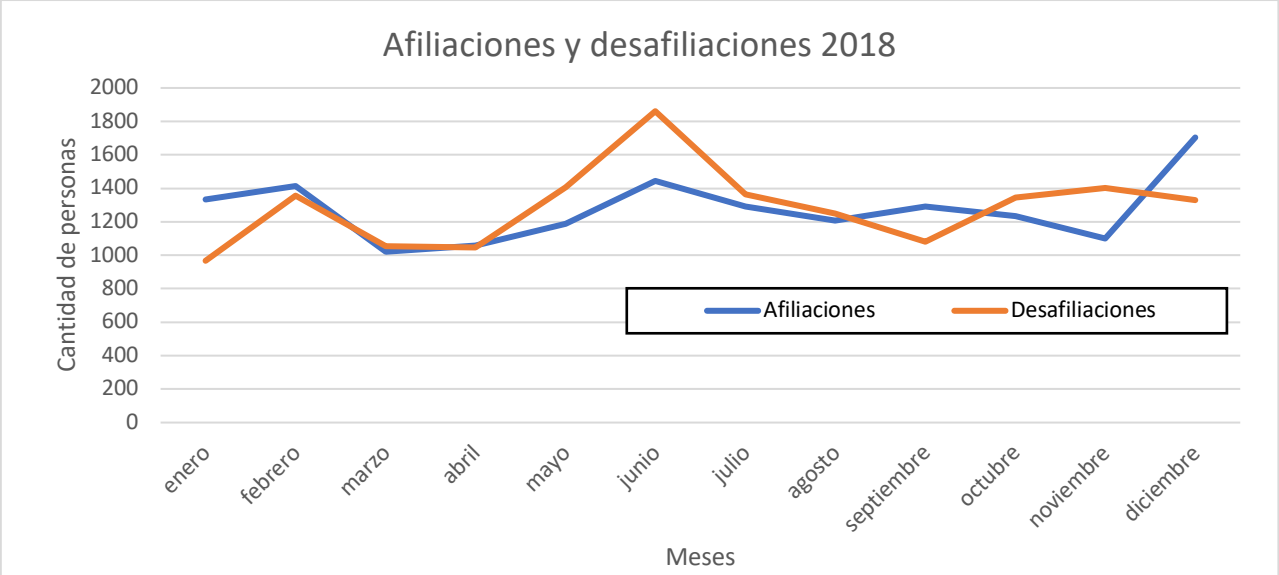
Dentro del la Dirección de Desarrollo, el proyecto se encuentra en el área de Evaluación de Proyectos, área en la que trabajan dos Ingenieros Civiles Industriales, un Ingeniero en Informática y un Ingeniero en Información y Control de Gestión. Constantemente se están recibiendo nuevas iniciativas de proyectos que se evalúan y desarrollan en esta área. Además, se generan los cruces de datos y análisis de información para una mejor toma de decisiones.

2.2. Situación sin proyecto

Como se mencionó anteriormente, el Programa Amigos es el medio por el cual Fundación las Rosas recibe mensualmente donaciones de los socios del programa, estos son del orden de 100.000 socios activos. Mensualmente personas se incriben en el programa y otras se desafilian, por lo que estas son

las dos primeras variables que pueden describir el desempeño del programa. El año 2017 se afiliaron 15.037 personas y se desafiliaron 15.047, por lo que se cerró ese año con 10 inscritos menos en el programa. Como se puede ver en el Gráfico 3, el año 2018 las afiliaciones y las desafiliaciones al programa fueron muy similares, no obstante, se cerró el año 2018 con 15.292 amigos nuevos y con 15.464 amigos fugados, esto es, 172 socios menos en el programa. Por lo anterior, se puede ver que el Programa Amigos se está reduciendo en cantidad.

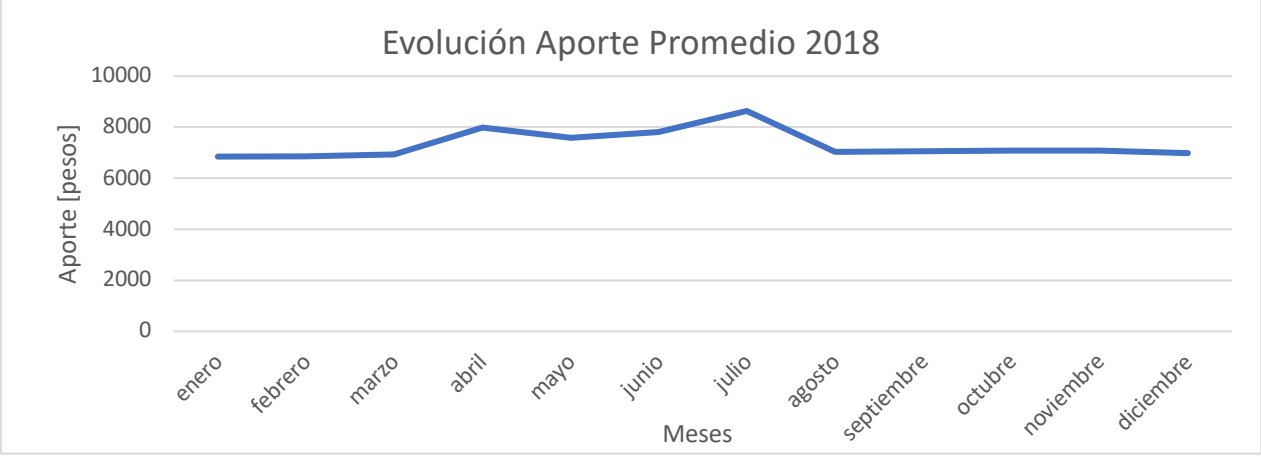
Gráfico 3 - Afiliaciones y desafiliaciones 2018



Fuente: Elaboración propia [5]

Por otra parte, es importante considerar es el aporte promedio. En el Gráfico 4 se puede ver que durante el año 2018 que se mantuvo relativamente constante, rondando en promedio los $7.000.

Gráfico 4 - Evolución Aporte Promedio 2018



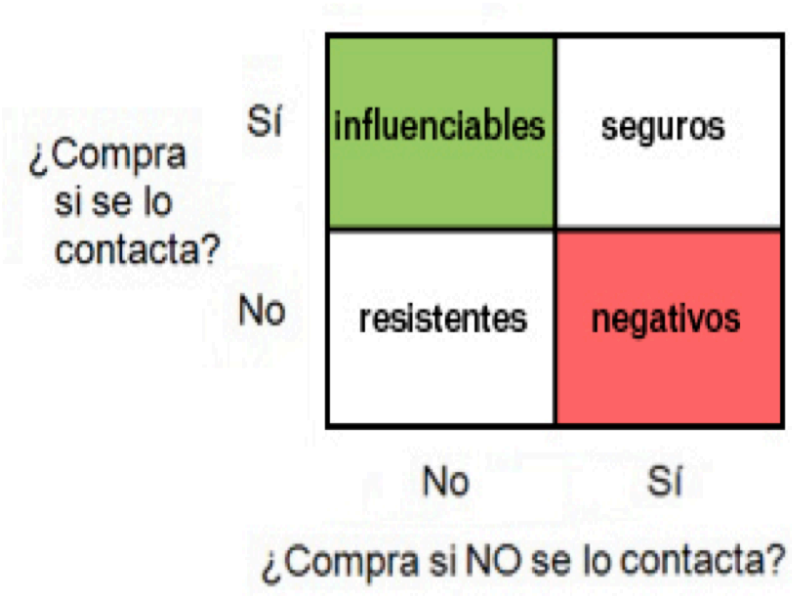
Fuente: Elaboración propia [5]

Considerando estas tres variables, se puede ver que el Programa Amigos no está creciendo como Fundación las Rosas necesita que crezca, esto porque los costos de la institución se están incrementando, por la disminución en la valencia de los residentes. Es así como, si se considera que el Programa Amigos aporta cerca del 40% de los ingresos, es que se puede suponer que puede ser crítico para la institución tener una reducción de ingresos por este ámbito. Así, si se sigue con la misma gestión, se finalizarán los próximos períodos con déficit.

2.3. Propuesta de mejora

Para explicar la propuesta de mejora, se presenta la segmentación fundamental propuesta por el modelamiento uplift (ver Ilustración 7), en la que se clasifican a los clientes en 4 categorías; influenciables, seguros, resistentes y no-molestar. Esta segmentación se basa en reconocer que los clientes son heterogéneos, por lo que actuarán de manera diferente frente a las campañas de marketing. Según esta segmentación, el único grupo que presenta una respuesta incremental positiva frente a los estímulos es el grupo de los influenciables, en los otros tres segmentos los estímulos son recursos perdidos o pueden generar una respuesta negativa de parte del cliente, pudiendo provocar la salida de la persona del programa. [24]

Ilustración 7 - Segmentación fundamental modelamiento uplift



Fuente: Marketing Directo – MAYSA consultores [35]

Además, el modelamiento uplift trabaja con un grupo de control, esto permite evaluar las campañas y crear indicadores de manera sincera, obteniendo el los beneficios de haber realizado la campaña, por sobre el no haber realizado ninguna acción. Con esto se pueden mejorar cada vez más las campañas, ya

que permite la realización de experimientos, así, entender el cmportamiento de los socios y rentabilizar de una manera más efectiva la cartera de donantes.

Como se mencionó anteriormente, el envío de correos de campañas se trabaja con envíos masivos a todos los afiliados al Programa Amigos, aquí es donde se propone el primer cambio, al poder reconocer a los donantes influenciables, se pueden focalizar las campañas y crear evaluaciones que midan el impacto genuino de esta, es decir, el valor agregado de la campañas.

Teóricamente, un modelamiento de este tipo mejora la tasa incremental de respuesta hasta 3 veces, disminuye los costos de las campañas en un 40% y reduce los efectos negativos al no contactar a los afiliados que no desean recibir la campaña.[25]

Para poder llevar a cabo este modelamiento, es necesario contar con información de las campañas, la que incluye el uso de un grupo de control. Como las campañas se han trabajado bajo toda la cartera de afiliados, el grupo de control es inexistente, por lo que se propone la realización de experimentos que permitan generar los datos necesarios para crear este modelo y poder evaluarlo. La principal ganancia está en el conocimiento que se puede generar para la institución, lo que permitiría enfocar las campañas y se traduciría en una baja de la tasa de fuga y aumento en la rentabilidad de las campañas.

Posteriormente, se podría extrapolar ese conocimiento al Contact Center, permitiendo enfocar las llamadas que realizan, haciendo más eficientre su trabajo, reduciendo horas hombre y generando los mismos beneficios para las campañas que ellos realizan. Esto claramente luego de evaluar si los socios se comportan de la misma manera en ambos canales, de no ser así, igualmente se puede generar un modelo tipo up-lift propio para el Contact Center, por medio de experimentos y una metodología similar. Finalmente, se puede comparar ambos modelos y analizar diferencias, similitudes y ver cómo se comportan.[5]

[5] Cabe destacar que todo acto con el Contact Center queda fuera del alcance de esta memoria, se incluyó en este último párrafo solo para mostrar nuevas posibilidades que se abren con la aplicación del modelo up-lift.

3. Objetivos

3.1. Objetivo General

Desarrollar diseño experimental que permita levantar información necesaria para aplicación de un modelo up-lift.

3.2. Objetivos Específicos

- Realizar análisis descriptivo que muestre cómo se comportan los socios del Programa Amigos para ver potenciales variables que influyen en que un socio acepte o no una campaña.
- Construir un modelo regresivo explicativo para obtener las combinación de variables idóneas para la realización de experimentos.
- Construir un diseño experimental.
- Proponer la forma de evaluar mediante un modelo tipo uplift.

4. Marco Conceptual

4.1. Sobre los modelos

En esta sección se habla sobre las herramientas utilizadas para la construcción y evaluación de los modelos

4.1.1. Estimación de los modelos

En este punto el estudio tiene dos niveles de investigación, por una parte, en un nivel explicativo, se utilizan regresiones lineales con el objetivo de encontrar causalidad en el comportamiento estudiado. Posteriormente, se busca encontrar un modelo que prediga y entregue una propensión a aceptar una campaña, por lo que se estará en un nivel de investigación predictivo. La principal diferencia entre estos es que el primero evalúa una hipótesis, mientras que el segundo realiza una estimación puntual.

4.1.1.1. Regresión Lineal

Un modelo de regresión lineal intenta explicar la relación entre una variable dependiente y un conjunto de variables independientes. La variable dependiente tiene la forma que se ve en la ecuación (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Donde,

$X_1, X_2, \ldots, X_n$ son las variables independientes,

$\beta_0, \beta_1, \ldots, \beta_n$ son los parámetros que miden la influencia de las variables independientes sobre Y,

ε es un factor de error que recoge la diferencia entre la predicción y el valor real.

a. Métricas de desempeño

La relación entre la variable dependiente y las independientes puede ser fuerte o débil. Para determinarlo se utilizan comunmente el coeficiente de relación de Pearson r, este coeficiente puede tomar valores en [-1,1], donde un valor 1 indica una relación positiva perfecta de las variables, una relación directa. Un valor de correlación -1 indica una correlación negativa perfecta, una relación inversa. Un valor de correlación 0 indica que las variables no están relacionadas, es decir, son independientes. Estos son los valores que se deben analizar para construir la regresión, ya que se deben seleccionar variables que estén correlacionadas con la variable dependiente, pero que no estén correlacionadas entre ellas.

El coeficiente de determinación múltiple R^2 indica la cantidad de varianza que el modelo logra explicar, este valor puede estar entre [0,1]. Un valor cercano a 0 indica bajo poder explicativo; mientras más cercano a 1 sea este valor, mejor es el poder predictivo del modelo. La significancia de cada variable dentro del modelo es evaluada con su *p-valor*. [33]

4.1.1.2. Regresión Logística

Los modelos de regresión logística (Logit) son del tipo regresión, esto es, que relaciona una variable dependiente con una o más variables independientes. Son utilizados cuando se tiene una elección discreta, muy útiles cuando se quiere modelar una probabilidad. El modelo viene dado por la fórmula:

$$\Pr(y = 1|x) = \frac{e^{\beta \cdot x}}{1 + e^{\beta \cdot x}}$$

Donde x es un vector de variables independientes, β es el vector de coeficientes de la regresión e y es la variable independiente que se quiere modelar. [26]

a. Métricas de desempeño

Para medir la eficiencia del logit se utilizan principalmente 4 herramientas, que son válidas para cualquier modelo de clasificación. Estas herramientas son: Matriz de confusión, la curva CAP y la curva ROC. La matriz de confusión compara la clase original de las observaciones con la clase predicha, ubicando todas las observaciones en 4 cuadrantes como se ve en la Tabla 1.

Tabla 1 - Matriz de confusión. Marco Conceptual

		Valor Real	
		1	0
Predicción	1	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	0	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Fuente: Elaboración Propia

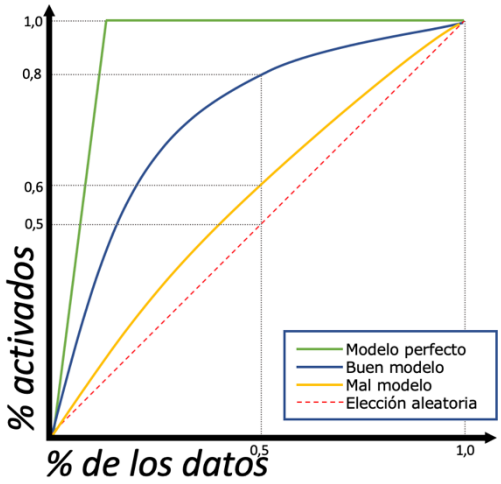
Existen 3 indicadores que se obtienen analizando esta tabla:

- *Recall o Sensitivity* → Cuánto detectó de todos los positivos. Entonces, el complemento 1-R: Cuánto positivo se está perdiendo. Se calcula como sigue: R: $\frac{VP}{VP+FN}$

- *Precision o Specificity* → Qué parte de lo que se predijo positivo, era efectivamente positivo. El completento, entonces, 1-*P*: A cuánto negativo se está tocando. Se calcula como sigue: P: $\frac{VP}{VP+FP}$

- Accuracy → Qué parte de los individuos está correctamente clasificada. Se calcula como sigue: A: $\frac{VP+VF}{Total}$

Ilustración 8 - Ejemplo curva CAP



Fuente: Elaboración propia

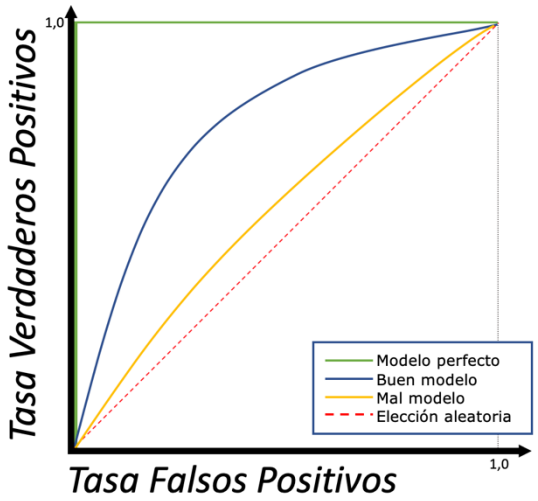
La curva CAP (Cumulative Accuracy Profile) ordena las observaciones de mayor a menor probabilidad de activación predicha. Se puede ver un ejemplo en la Ilustración 8, en la que se muestran 4 curvas, la curva en rojo representa un modelo que clasifica aleatoriamente, así, en un 50% de los datos, tenemos un 50% de los activados. Un modelo como el graficado con amarillo indica que lo puede ser un modelo con poco poder de clasificación o ruido, por lo que no se debiese considerar.

Un modelo como el que aparece en azul se considera un buen modelo, notar que está muy por encima de la curva aleatoria y que al 50% de los datos ya ha encontrado al 80% de la clase activada realmente. Finalmente el modelo perfecto aparece en verde, en un comienzo encuentra todas las observaciones de la clase activada.
La curva ROC, por su parte, grafica cómo se comporta el *Recall* versus el 1-*Specificity*, también llamada la tasa de falsos positivos. Aquí se puede analizar el AUC, el área bajo la curva (*area under curve*). En la Ilustración 9, se puede ver en verde el modelo perfecto que tiene un área igual a 1, ya que llegó al 100% de los verdaderos positivos sin caer en ningún falso positivo.

Análogo al anterior, también se ve en rojo la curva de un modelo aleatorio, esta curva tiene un AUC de 0,5.

Finalmente las curvas azul y amarillo que representan un buen y un mal modelo respectivamente. Valores de AUC cercanos al 0,6 se consideran ruido, lo que en la imagen sería la curva amarilla. Valores de AUC por sobre los 0,7 ya se considera bueno y que el modelo tiene capacidad predictiva, lo que en este caso correspondería a la curva azul.

Ilustración 9 - Ejemplo curva ROC



Fuente: Elaboración propia

4.1.1.3. Random Forest

Random Forest (RF) es una técnica de ensamblado de Machine Learning basa en el principio *The Wisdom of Crowds* (la sabiduría de los grupos). Esto porque todos los árboles actúan como un comité en el que se decide la clase resultante (1 ó 0 en este caso) de cada observación.

Esta técnica funciona porque los árboles se protegen entre ellos de sus errores individuales, es decir, sus errores no están correlacionados, debido a técnicas empleadas en su construcción. La primera de ellas es el *Bagging* o Bootstrap Aggregating, consiste en que cada árbol trabaje con una porción de la base, seleccionada aleatoriamente, con reemplazo. Además, cada árbol trabaja con un set de variables dependientes, que se selecciona aleatoriamente. Estas dos técnicas, en conjunto, provocan que los árboles tengan mayor variabilidad.

a. Balanceo de clases

Al utilizar sólo una porción de observaciones de la base (*Bagging*), aparecen problemas cuando no se tienen bases balanceadas, esto es, que las cantidades de observaciones de cada clase son muy diferentes, lo que hace que el modelo resultante esté más apegado a la clase mayoritaria que lo que debería. Por ejemplo, si tengo una base con 9M observaciones de clase 1 y 1M observaciones de clase 0, si cada árbol usa 1M observaciones, habrán muchos árboles que tengan pocas o no tengan ninguna observación de clase 0, por lo que esos árboles (casi) siempre votarán hacia la clase 1, la clase mayoritaria.

Para evitar lo anterior, se debe balancear la base, eligiendo aleatoriamente en las bases. Se puede emplear *up-sampling* de la clase minoritaria, esto es,

elegir, aleatoriamente con reemplazo, N observaciones de una base con tamaño M, con N>M. La otra alternativa es mediante *down-sampling* de la clase mayoritaria, esto es, elegir aleatoriamente, sin reemplazo, N observaciones de una base con tamaño M, con N<M.

Además, se puede trabajar con distintos niveles de balanceo, por ejemplo, 25%, 50%, 90%, de la clase mayoritaria. Jugar con este valor es de gran utilidad para obtener conclusiones y analizar los distintos ajustes obtenidos de distintos niveles de balanceo.

b. Métricas de desempeño

Random Forest puede utilizar las métricas de desempeño desarrolladas anteriormente para el Logit, esto ya que es un modelo de clasificación. Además, RF cuenta con los *Out Of Bag error*(OOB error), un métrica que sirve para las técnicas que utilizan *Bagging*.

Esto ya que, al utilizar *Bagging*, cada árbol es entrenado con un subset de observaciones, esto implica que, para cada observación, hay un conjunto de árboles que no fueron entrenados con esa observación. La clase predicha, por este conjunto de árboles, que no vieron la observación en el entrenamiento, es lo que se utiliza para calcular el OOB. Entonces, la medida OOB es la porción de observaciones correctamente predicha por el conjunto de árboles anteriormente definido. Cabe destacar que para cada observación de la base hay un conjunto de árboles diferente.

Así, el OOB error es una medida de error interna de un RF, que se calcula mientras este se construye. La principal ventaja es que puede dar una idea de qué tan bien lo está haciendo el modelo sin tener que contar con un set de prueba, pero tiene a subestimar el error.

c. Hiperparámetros

Existen 5 parámetros que determinarán el desempeño del modelo, usualmente, los programas computacionales tienen valores predefinidos para estos parámetros que generan buenos resultados, pero se pueden ajustar dependiendo de los objetivos de la investigación. Estos parámetros son:

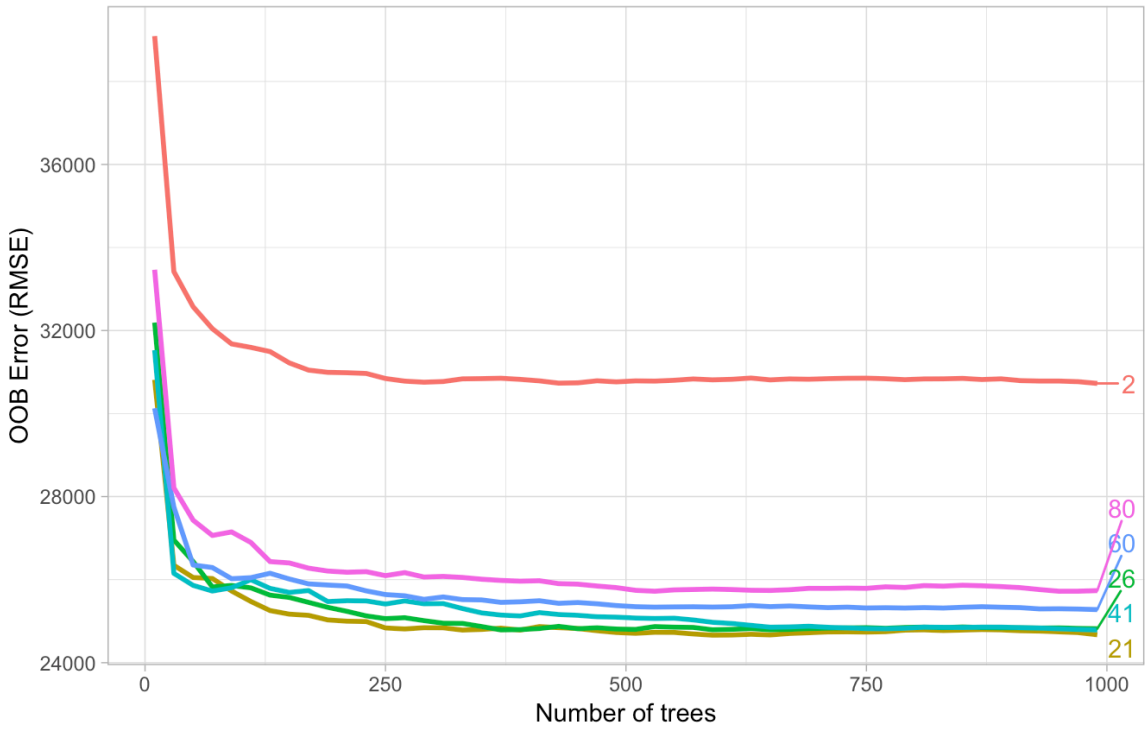
- Cantidad de árboles
- Cantidad de variables por árbol
- Tamaño de los nodos
- Porción de la base a utilizar
- Criterio de corte

La cantidad de árboles debe ser lo suficientemente grande para que el error esté estable, teniendo consideración en que agregar más árboles ya no mejorará el modelo y representa un costo a nivel computacional.

La cantidad de variables de cada árbol dependerá de qué cuántas variables relevantes se tengan, si hay pocas variables relevantes un mayor valor de este parámetro tenderá a tener un mejor resultado, si hay muchas variables relevantes, es mejor tener menos variables. De los 5 parámetros, este es el que mayor influencia tiene en la precisión del modelo.

A modo de ejemplo, en la Ilustración 10 se puede ver un gráfico que muestra cuánto se equivoca (*OOB Error)* el modelo de Random Forest según la cantidad de árboles. Se puede ver que con una menor cantidad de árboles el error va aumentando, y a medida que se aumenta este parámetro el error se va estabilizando. Las distintas curvas en el gráfico corresponden a distintos valores para la cantidad de variables por árbol, estando las cantidades al lado derecho de la imagen.

Ilustración 10 - OOB Error vs Cantidad de árboles. Ejemplo



Fuente: Hands-on Machine Learning with R, Chapter 11 Random Forest [34]

Los restantes parámetros, el tamaño de los nodos, la porción de la base a utilizar y el criterio de corte; son parámetros que pueden tener alguna influencia en la precisión del modelo, pero esta es baja en comparación con los primeros. Son comunmente utilizados para mejorar eficiencias computacionales, en modelos muy complejos. Al no ser el objetivo de esta memoria llegar a un modelo muy robusto y eficiente, no se profundizará en ellos. [34]

4.1.2. Clustering

El *clustering* o agrupamiento es una técnica usada en minería de datos que consiste en agrupar datos de acuerdo a algún criterio de distancia o similitud. Esto ayuda en las construcciones de modelos ya que permite obtener conclusiones a nivel agregado, es decir, poder crear categorías y concluir respecto a ellas.

A la hora de formar clusters, se está asumiendo que existen efectivamente esos grupos y que se pueden representar de acuerdo a las variables observables que se tiene. Existen métodos para evaluar la precisión de los clusters, estos están basados en la distancia entre los puntos y el centroide de cada cluster, pueden ser útiles, pero finalmente, si se utiliza o no cada cluster, va a depender de la utilidad que este presente para el modelo, lo que se ve reflejado en la significancia individual y en cuánto mejora (si es que lo hace) la precisión del modelo completo.

También es importante notar la utilidad del cluster, esto es, que efectivamente cree distintas categorías, que logre separar los datos, ya que si un cluster no logra agrupar los datos de una manera en la que se puedan sacar conclusiones, no tendría sentido su aplicación. Por ejemplo un grupo que contenga el 5% de los individuos no se puede considerar como una categoría si es que se desea un número reducido de categorías, ya que quedarían desbalanceadas en cantidad de personas, es decir, no es lo suficientemente representativo.

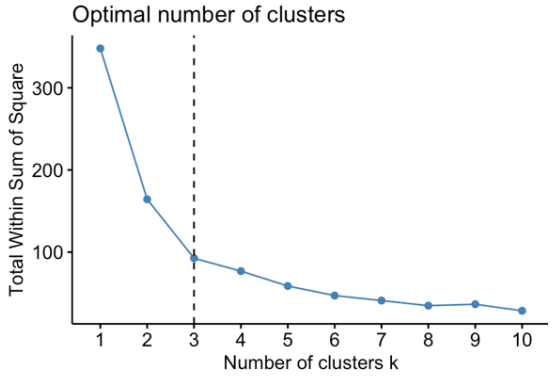
- **K-means**

Kmeans es un método de partición que asigna los individuos a K grupos, esto minimizando todas las distancias cluster-individuo y maximizando la distanticia cluster-cluster.

Este método se escogió ya que es muy popular en la literatura por su simpleza, requiere poco procesamiento computacional y llega rápidamente a óptimos. No obstante, requiere que se le especifique previamente el número de grupos a formar, para lo que existen distintos criterios.

El método *del codo* es idóneo para trabajar con K-means, ya que grafica la suma cuadrada de las distancias individuo-cluster (valor que K-means trata de minimizar) según la cantidad de clusters (ver Ilustración 11). Así, recomienda que se agreguen clusters hasta que agregar uno extra no signifique una baja considerable en esta suma. Se llama así ya que la curva del gráfico tiene un quiebre pronunciado (el codo) en ese lugar. En el gráfico de ejemplo, la cantidad recomendada por este método son 3 clusters.

Ilustración 11 - Ejemplo cantidad de clusters



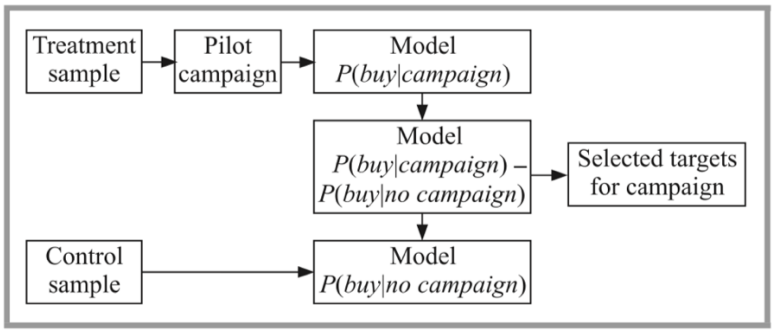
Fuente: Elaboración propia

4.1.3. Modelamiento Uplift

Los modelos uplift son del tipo incremental, esto es, que están diseñados para explicar el aumento en una conducta, causada por un estímulo. En marketing se utilizan comunmente para focalizar campañas, ya que busca el segmento de clientes que responderá positivamente al estímulo, dejando fuera de la campaña a los que no tendrán respuesta, y a los que tendrán una respuesta negativa.

Para clasificar a los clientes, los modelos uplift se basan en la segmentación fundamental explicada en el apartado 2.3. Como se explicó anteriormente, estos modelos requieren de un grupo de control ya que, al medir el efecto en el grupo que recibió el estímulo, y restarle las respuestas del grupo control, que no recibió estímulo alguno, se obtiene una medida de la diferencia entre aplicar o no el estímulo, una valoración genuina de las campañas de marketing.
Existen dos maneras de crear los modelos uplift, la más evidente es construir dos modelos, uno para el grupo de tratamiento y otro para el grupo de control, luego restarlos (ver Ilustración 12). El problema de este método es que se concentra mucho en el modelamiento de cada uno de los grupos en lugar de enfocarse en la diferencia entre ellos, lo que provoca que el modelo sea menos representativo de la realidad, ya que la variación de la diferencia es menor que las variaciones en cada modelo por separado.[25]

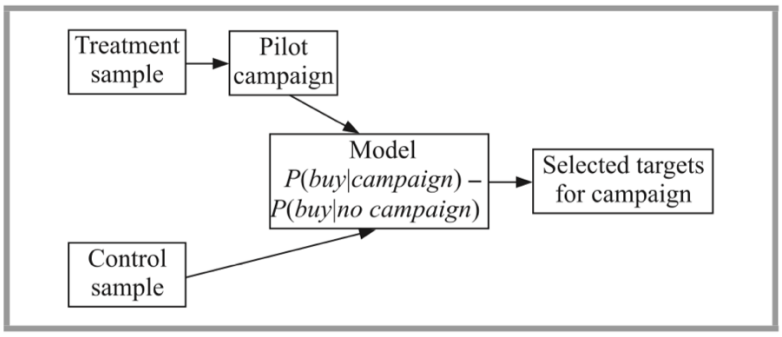
Ilustración 12 - Modelamiento uplift desde dos clasificadores



Fuente: Rzepakwski and Jaroszewicz, (2012). [25]

El otro enfoque es modelar directamente la diferencia entre ambos grupos (ver Ilustración 13). Estos modelos buscan maximizar la diferencia entre ambos grupos basándose en distintos criterios de divergencia.

Ilustración 13 - Modelamiento uplift directo



Fuente: Rzepakwski and Jaroszewicz, (2012). [25]

Comúnmente se utilizan dos medidas de divergencia, Kullback Leiber (KL) y la Distancia Euclidiana (E)[40]. Esas medidas de divergencias, desde una distribución $Q = (q_1, \dots, q_n)$ a una distribución $P = (p_1, \dots, p_n)$ se definen de la siguiente manera:

$$KL(P:Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

$$E(P:Q) = \sum_i (p_i - q_i)^2$$

4.2. Evaluación de Campañas

4.2.1. Tasa de apertura

Corresponde al porcentaje de correos abiertos de una campaña. Se calcula de la siguiente manera:

$$TA = \frac{Total\ de\ correos\ abiertos}{Total\ de\ correos\ enviados}$$

4.2.2. Tasa de respuesta

Corresponde al porcentaje de destinatarios que abrieron el correo y además completaron la acción deseada. Se calcula de la siguiente manera:

$$TA = \frac{Total\ de\ conversiones}{Total\ de\ correos\ enviados}$$

4.3. Diseño de Experimentos

Los experimentos se diseñan con el objetivo de encontrar si existen diferencias significativas entre 2 poblaciones: un grupo de tratamiento y un grupo de control, cada uno de estos debe tener un tamaño mínimo para que la conclusión tenga validez.

4.3.1. Grupo de tratamiento y grupo de control

Con el objetivo de encontrar el verdadero efecto de una acción, se trabaja con dos grupos de personas. Al grupo de tratamiento se le aplica la acción a medir, al grupo de control no se le aplica. La diferencia entre los resultados de ambos grupos es el resultado atribuible a la acción.

Es importante considerar que para formar estos grupos se seleccionan clientes aleatoriamente, además se debe evaluar si ambos son representativos de la población total en cuanto a variables demográficas, por ejemplo cantidad de hombres y mujeres, distribuciones de edad, zonas de residencia, entre otras.

4.3.2. Planteamiento de hipótesis

La diferencia que se quiere estudiar puede estar en base a si se quieren comparar medias o proporciones y se debe plantear en forma de hipótesis. El caso típico es plantear la hipótesis nula como $H_0: \alpha_1 = \alpha_2$, que se puede aceptar o rechazar dependiendo del valor del estadístico que se calcula y el intervalo de confianza.

En caso de rechazar H_0, se sabe que los valores son distintos, lo que puede tener 3 implicancias diferentes, que son los valores que puede tener la hipótesis alternativa H_A.

$$H_A^1: \alpha_1 < \alpha_2$$
$$H_A^2: \alpha_1 > \alpha_2$$
$$H_A^3: \alpha_1 \neq \alpha_2$$

El valor de H_A depende de si la hipótesis es unilateral o bilateral, en el primer caso se considera que uno de los parámetros debe ser mayor que el otro, mientras que en el segundo, cualquiera de los dos parámetros puede ser mayor. Las hipótesis bilaterales son más conservadoras y disminuyen el riesgo de cometer el error tipo I.

4.3.3. Tamaños muestrales

Para el cálculo de tamaños muestrales se requieren de los siguientes datos:

- Magnitud de la diferencia.
- Una idea aproximada de los parámetros de la variable que se estudia.
- Seguridad del estudio (riesgo de cometer error tipo I) → Z_α
- Poder estadístico (riesgo de cometer error tipo II) → Z_β
- Definir si la hipótesis será unilateral o bilateral, esto afectará en el valor final de Z_α.

Para el caso donde se quieran comparar dos proporciones, los tamaños muestrales se calculan de la siguiente manera:

$$n = \frac{\left[Z_\alpha \cdot \sqrt{2\bar{P}(1-\bar{P})} + Z_\beta \cdot \sqrt{P_1(1-P_1) + P_2(1-P_2)}\right]^2}{(P_1 - P_2)^2}$$

Donde P_1 es la proporción en el grupo de control, P_2 es la proporción en el grupo de tratamiento y $\bar{P}$ corresponde a la media entre P_1 y P_2.

En casos donde se quiere comparar dos medias, los tamaños muestrales se calculan de la siguiente manera:

$$n = \frac{2\left(Z_\alpha - Z_\beta\right)^2 \cdot S^2}{d^2}$$

Donde S^2 es la varianza en el grupo de control y d es el valor mínimo de la diferencia que se desea testear.

Además, se debe hacer un ajuste por las pérdidas esperadas, esto provocado por correos mal digitados, rebotados, entre otras causas. Para una proporción R de pérdidas esperadas, el tamaño de la muestra se debe multiplicar por $\frac{1}{1-R}$.

4.3.4. Cálculo de estadísticos

4.3.4.1. Test de Proporciones

Si para las poblaciones 1 y 2, se tienen de proporciones P_1 y P_2 respectivamente, la hipótesis nula sería:

$$H_0: P_1 = P_2$$

Así, para este test se debe calcular el estadístico z de la siguiente manera:

$$z = \frac{P_1 - P_2}{EED} = \frac{P_1 - P_2}{\sqrt{P \cdot (1-P) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Donde EED corresponde al error estándar de la diferencia, se utiliza una aproximación con un valor de P mezclado. Una buena aproximación aproximación para P es la que se encuentra más abajo. Los valores de n_1 y n_2 son los tamaños de las poblaciones 1 y 2. [22 y 23]

$$P \cong \frac{n_1 \cdot P_1 + n_2 \cdot P_2}{n_1 + n_2}$$

4.3.4.2. Test de Medias

Si para las poblaciones 1 y 2 se tienen las medias μ_1 y μ_2 respectivamente, la hipótesis nula sería:

$$H_0: \mu_1 = \mu_2$$

Así, para este test se debe calcular el estadístico z de la siguiente manera:

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

Donde σ_i representas las desviaciones estándar de las poblaciones 1 y 2, y los valores de n_1 y n_2 son los tamaños de las poblaciones 1 y 2. [22 y 23]

5. Metodología

En general, en los proyectos de minería de datos se trabajan dos metodologías: KDD y CRISP-DM. La gran diferencia es que CRISP-DM incluye al inicio una fase de entendimiento del negocio. Se considera que esta fase está resuelta por los apartados 1, 2 y 3, por lo que se seguirá con la metodología KDD. [20, 21 y 28]

5.1. KDD

La metodología KKD (Knowledge Discovery in Databases) implica el la evaluación e interpretación de patrones para tomar decisiones, con el objetivo de crear conocimiento desde la información. Trabaja con 5 fases: Selección, Procesamiento, Transformación, Minería de Datos e Interpretación/Evaluación (ver Ilustración 14).

Ilustración 14 – KDD

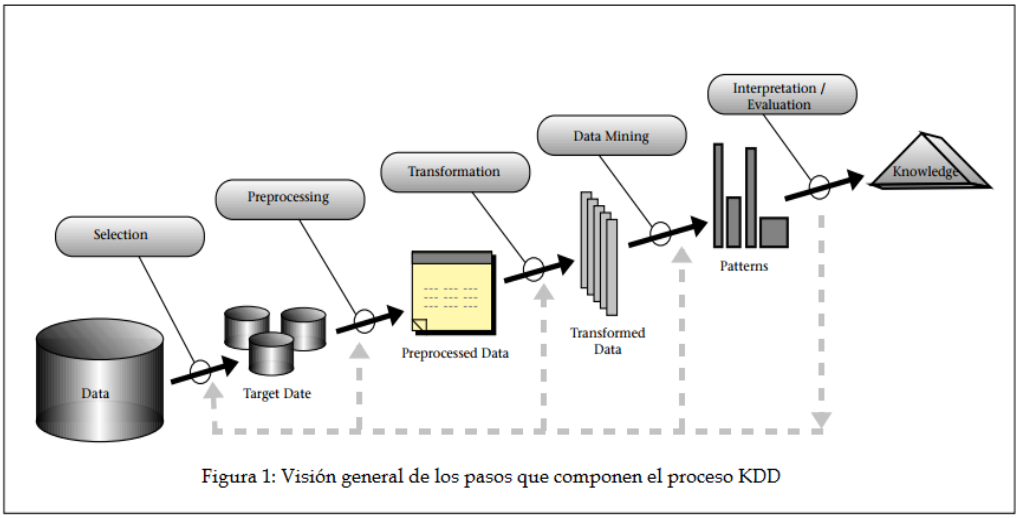


Figura 1: Visión general de los pasos que componen el proceso KDD

Fuente: Descripción funcional – CareCloudExtern [30]

a. Selección

Reconocimiento de las bases de información, creación de objetivos de la investigación

b. Procesamiento

En esta estapa ocurre la impieza de la data, transformación de variables, eliminación de ruido, valores atípicos o inconsistentes. Se debe decidir qué hacer con los datos faltantes, comunmente se trabajan dos opciones, una es eliminar las observaciones con datos faltantes, otra es completar con la moda o la mediana.

Ocurre tambien la integración de las bases, en caso de que vengan con distintas fuentes de información.

c. Transformación

Se estructura la data de la mejor manera para las herramientas que se quieren utilizar en la construcción de los modelos.

Se realiza un análisis descriptivo con el objetivo de encontrar los primeros patrones, y formarse una primera idea o una intución de cómo debiese ser el modelo.

d. Minería de Datos

Se construyen los modelos en un proceso iterativo, en esta etapa es muy probable tener que volver hacia atrás para la creación de nuevas variables de interés, limpieza de alguna variable o alguna transformación que no se haya realizado previamente.

e. Interpretación/Evaluación

La interpretación de estos modelos y la obtención de conocimiento va a alimentar principalmente el diseño de experimentos.

5.2. Diseño e Implementación de Experimentos

En esta etapa se deben establecer los objetivos de los experimentos, qué es lo que se desea testear, con qué factores y en qué niveles se desea trabajar. Esto debe ser definido, idealmente, en conjunto con la institución.

Posteriormente, se deben definir las hipótesis a testear, estudiar la factibilidad, crear las campañas de los experimentos y definir los grupos muestrales y evaluar económicamente el experimento. Finalmente se procede con la implementación con el envío de las campañas.

- **Evaluación de campañas y experimentos**

En esta etapa se estudian los resultados de los experimentos, se debe comparar los resultados del grupo de tratamiento y de control mediante el test z de proporciones o de medias según cómo se diseñó el experimento. Aquí se aceptan o rechazan las hipótesis.

Una vez aceptada o rechazada la hipótesis nula se debe concluir con respecto a la implicancia económica y efectividad de la campaña.

5.3. Confección Modelo Uplift

En esta etapa se debe utilizar nuevamente KDD para la confección de un modelo uplift, de acuerdo a lo encontrado en los experimentos, se procede con la construcción y evaluación del modelo uplift. Se busca poder clasificar a los donantes según la segmentación fundamental.

Para ello, se calculará el lift de cada persona, de acuerdo a lo encontrado en los experimentos, posteriormente se ordenaran de mayor a menor Lift y se procederá analizar los distintos deciles.

5.4. Conclusiones y Recomendaciones

Finalmente, se realizan conclusiones y recomendaciones a partir de lo encontrado en la investigación. Se busca mostrar el conocimiento generado para que pueda ser comprendido por la institución, se deben elaborar recomendaciones.

Además, se debe evaluar qué tan replicable es la metodología o la por parte de la institución.

6. Alcances

El proyecto presentado trabaja con las campañas de Fundación las Rosas. Se trabajará y analizará la información histórica de las campañas anteriores que se posee, estas corresponden a "7 lucas del confort", venta de SOAP[6] 2018, venta de SOAP 2019 y Navidad 2018

Con respecto a los datos, se trabajará con los datos que estén digitalizados, se procesarán y estructurará la base para el estudio y construcción de los modelos, pero no se considera realizar cambios en cómo Fundación las Rosas almacena su información ni la estructura los datos para las aplicaciones futuras y calibraciones del modelo.

Se realizará el diseño de los experimentos con los que se quiere levantar la información necesaria para la construcción del modelo y la generación de conocimiento, pero no se considera la implementación, esto ya que los tiempos de la Fundación no calzan con los de la memoria.

[6] Seguro obligatorio de accidentes personales

7. Desarrollo Metodológico

7.1. Selección

En esta etapa se reconocieron las bases de datos. La primera contiene la información demográfica de los Amigos, la segunda contiene información de transacciones y hay un tercer grupo de base pequeñas y aisladas con la información de las campañas.

Se poseen del orden de 130 mil observaciones en la base de datos demográficos y del orden de 3 millones de observaciones en la base de datos transaccionales.

7.2. Procesamiento y transformación

7.2.1. Base de Datos Demográficos

En esta base se encuetra la información de los socios del Programa Amigos. Las variables presentes inicialmente en esta base están en la Tabla 2.

Tabla 2 - Variables originales. Base demográfica

Nombre Variable	Descripción	Tipo
id	Número identificador	Entera
Genero	[M, F, E] Masculino, femenino o empresa	Categórica
Region	[0,15] Región a la que pertenece	Categórica
ProvinciaCiudad	54 posibles valores. Provincia a la que pertenece	Categórica
Comuna	321 posibles valores. Comuna del amigo	Categórica
ZonaRegion	8 posibles valores. Región del amigo si FLR tiene presencia. "Otras" si no	Categórica
AportePesos	Monto de aporte en pesos	Entera
MesesAporte	Periodos en los que ha efectivamente aportado	Entera
CodRecauda	Código institución bancaria	Categórica
NomRecauda	Nombre institución bancaria	Categórica
TipoRecauda	[1, 4] Código tipo de pago	Categórica
TipoRecaudaNom	[Bancos, Domicilio, Tiendas Comerciales, Transbank] Nombre tipo de pago	Categórica
Frecuencia	[Mensual, Bimensual, Semestral, Anual]	Categórica
FecIngreso	Fecha de ingreso al programa	Entera
FecNace	Fecha de nacimiento	Entera

Fuente: Elaboración propia

La primera limpieza que se hizo fue eliminar las observaciones que tuvieran valor "Empresa" en la variable *Genero*, esto ya que se consideró que las empresas funcionan de manera distinta en el proceso de tomas de decisiones, y en la gestión del programa y comunicaciones, además de consideró que no existían la cantidad suficientes para tomas decisiones estadísticas al ser inferiores del 5%.

Por otra parte, se eliminó a las personas que no tienen frecuencia de aporte Mensual, esto ya que no representan una porción significativa de la población (menor al 5%), y para el cálculo de nuevas variables van a ensuciar la muestra por no tener pagos en todos los periodos. Finalmente, se eliminaron a las personas que tienen un medio de pago de Tiendas Comerciales y Domicilio, esto, por lo mismo que lo anterior, no son representativas de la población y funcionan de manera diferente, dejando sólo a los que tienen el valor "Bancos" que equivale a una cuenta corriente y "Transbank" que equivale a una tarjeta de crédito.

Posteriormente, se crearon nuevas variables a partir de las existentes. Con la fecha de nacimiento e ingreso se calcularon la edad actual (*Edadhoy*) y la edad de ingreso al programa (*Edadin*). Además se crearon dos variables para indicar el intervalo en que están las edades anteriores (*IntervaloEdadHoy* e *IntervaloEdadIn).* Los rangos de los intervalos son: [18, 34], [35, 64] y [65+].

Con la fecha de ingreso se calculó la cantidad de periodos que ha estado la persona dentro del programa(*MesesAdentro*), luego se calculó los periodos que no ha aportado y un índice de incumplimiento histórico (*incumplhist*) de la siguiente manera.

$$MesesSinPago = MesesAdentro - MesesAporte$$
$$incumplhist = MesesSinPago/MesesAdentro$$

De un análisis de percentiles (ver Ilustración 18 en 65) de la variable *AportePesos* se encontró que el 99% de los amigos tienen donaciones iguales o inferiores a $26.623. En el 1% restante, hay Amigos que donan hasta $500.000 mensuales. Dado esto, se separó la base en 3 rangos de aportes, el primero con montos de aportes pertenecientes a $[0, 30.000]$, el segundo con aportes en el rango $[30.001,\ 100.000]$, finalmente el último con aportes de $[100.001, 500.000]$. Las cantidades de amigos en cada rango son 128.903, 570 y 32 respectivamente, dadas estas cantidades, solo se construirá un modelo matemático para el primer rango, ya que los restantes no tienen una cantidad suficiente para alimentar estos modelos, no obstante, se incluye un análisis descriptivo de estos rangos.

Finalmente, se eliminaron las observaciones con datos faltantes, principalmente fueron por ausencia de fechas de nacimiento (y por ende la

edad). Además se eliminaron los datos inconsistentes y extremos, aquí se eliminaron amigos con edades negativas y menores a 18.

7.2.2. Base de Datos Transaccionales

En esta base se encuentran las donaciones de los años 2017 y 2018 completos, además desde enero hasta abril de 2019 de todos los Amigos, se tiene el monto donado por cada período, con la fecha exacta en la que se realizó el cargo. Las variables de esta base aparecen en la Tabla 3.

Tabla 3 - Variables originales. Base transaccional

Nombre Variable	Descripción	Tipo
id	Número identificador	Entera
FechaAporte	Fecha exacta en la que se realizó el cargo	Entera
PeriodoAporte	Periodo al que corresponde el cargo, de la forma AAAA-MM	Entera
MontoAporte	Monto aportado en pesos	Entera

Fuente: Elaboración propia

La limpieza de esta base consistió en asignar correctamente los periodos cancelados. Esto ya que muchas personas presentaban pagos dobles en algún periodo y una laguna en el periodo anterior o posterior.

A modo de ejemplo, se muestra un caso real. Se tenían los siguientes datos expuestos en la Tabla 4:

Tabla 4 - Ejemplo limpieza de datos, antes

id	FechaAporte	PeriodoAporte	MontoAporte
3601	30/04/18	2018-04	$5.000
3601	31/05/18	2018-05	$5.000
3601	29/06/18	2018-06	$5.000
3601	31/07/18	2018-06	$5.000
3601	31/08/18	2018-08	$5.000

Fuente: Elaboración propia

Se corrigió como se muestra en la Tabla 5:

Tabla 5 - Ejemplo limpieza de datos, después

id	FechaAporte	PeriodoAporte	MontoAporte
3601	30/04/18	2018-04	$5.000
3601	31/05/18	2018-05	$5.000
3601	29/06/18	2018-06	$5.000
3601	31/07/18	2018-07	$5.000
3601	31/08/18	2018-08	$5.000

Fuente: Elaboración propia

Existió la intención de obtener algún tipo de información a partir de la fecha exacta, no obstante esto se descartó, ya que, por lo declarado por la institución, el día de pago depende de la institución bancaria, no del socio, por lo que la información que se pudo haber obtenido hubiese sido producto de la gestión del programa y no de comportamiento de los socios.

Una vez corregidos los periodos de aporte, se traspuso la base para obtener un panel de datos, teniendo la información de cada persona en una fila. El panel resultante es de la forma como se muestra en la Tabla 6.

Tabla 6 - Ejemplo panel resultante

id	monto.201701	monto.201702	monto.201703	…	monto.201904
1877	$5.000	$5.000	$5.000	…	$8.000
1881	$10.000	NA	$10.000	…	$10.000

Fuente: Elaboración propia

Con este panel se crearon las variables de interés para el análisis. En primera instancia se creo una variable que cuenta la cantidad de aumentos que tiene cada persona, esto es, cuántas veces el monto de un período fue mayor al monto anterior en al menos $500. Se obtuvo que 14.745 personas tuvieron un aumento, 373 dos aumentos y 2 tres aumentos. Se considera que las personas con más de un aumento no son suficientes como para crear una variable que contenga esta información. Para efectos futuros, para las personas con más de un aumento se les consideró el último aumento.

Se desea obtener la información de la persona justo antes de hacer el aumento, así, se creó la variable *AportePrevio* que contiene el monto que aportaba el socio antes del aumento y se corrigió la variable *MesesAporte* según cuando fue el aporte.

Además, se crearon dos variables para utilizar como dependientes: *porc_ult_aum* que contiene el porcentaje de aumento, calculado como se muestra más abajo. La otra variable en *binaria*, que asigna un valor 1 si la persona hizo un aumento en la ventana de estudio y un 0 si no.

$$P_{ultaum} = \frac{AporteActual - AportePrevio}{AportePrevio}$$

7.2.3. Información de las campañas

Se tiene la información de 4 campañas, dos de ellas realizadas en marzo de 2018 y marzo de 2019 para la venta de SOAP. En estas campañas participaron 2.561 y 2.228 personas respectivamente.

En agosto de 2018 se realizó una campaña que consistió en solicitar la donación de los $7.000 de compensación por la colusión del papel higiénico. En esta campaña participaron 3.042 personas.

Finalmente se tiene la información de la campaña de Navidad de 2018, en este caso se cuenta con 505 socios participando. Por la baja participación de esta campaña no se consideró en el estudio ya que puede ensuciar más de lo que aporta.

Con las tres campañas a utilizar se crearon 3 variables binarias que indican si la persona participó o no en la campaña. Además, se creo una variable *participa* que corresponde a la suma de estas tres binarias.

7.2.4. Base resultante

Se integraron las bases, luego, con *Provincia* de la primera base y *AportePrevio* de la segunda, se creó una variable que indica cual es proporción de *AportePrevio* con respecto a un promedio de aporte de la provincia, el cálculo se hizo como *AportePrevio/PromedioProvincia*. Para el cálculo del promedio se escogió la Provincia por sobre la Comuna y Región porque presentaba el nivel de agregación indicado, la Región agrupa muchas personas diferentes y en Comuna habían casos en que eran muy pocos individuos para calcular un promedio representativo. Sin embargo, existen comunas con más de 1000 socios en los que se puede calcular un promedio, en estos casos se optó por el promedio de la comuna.

Finalmente, se reescalaron las variables, esto por las dimensiones en las que se encuentran, por ejemplo *AportePrevio* contiene valores de pesos, por lo que tiene un máximo valor de 30.000, esto, en comparación con *incumplhist* que tiene un máximo de 1 hace que los resultados de los modelos no permitan comparar el efecto de cada una de las variables. Por lo que, para el reescalado, a cada valor se le restó el mínimo y posteriormente se dividió por el nuevo máximo (máximo_previo – mínimo). Esto deja las variables numéricas con valores entre 0 y 1.

La base resultante contiene las variables que se muestran en la Tabla 7, las celdas en blanco indican que no hay cambio con respecto a las tablas anteriores, esto para destacar dónde ocurrieron las transformaciones de la base.

Tabla 7 - Base resultante

Nombre Variable	Descripción	Tipo
id	Número identificador	Entera
Genero	[M, F] Masculino o Femenino	Categórica
Region	[0,15] Región a la que pertenece	Categórica

ProvinciaCiudad	54 posibles valores. Provincia a la que pertenece	Categórica
Comuna	321 posibles valores. Comuna a la que pertenece	Categórica
ZonaRegion	8 posibles valores. Región del amigo si FLR tiene presencia. "Otras" si no.	Categórica
AportePesos	Monto de aporte actual	Entera
MesesAporte	Corregida al periodo previo al aumento. Reescalada	Numérica
CodRecauda	Código institución bancaria	Categórica
NomRecauda	Nombre institución bancaria	Categórica
TipoRecauda	[3, 4] Código tipo de pago	Categórica
TipoRecaudaNom	[Bancos, Transbank]	Categórica
Frecuencia	[Mensual]	Categórica
FecIngreso	Fecha de ingreso al programa	Entera
FecNace	Fecha de nacimiento	Entera
EdadIn	Edad de ingreso al programa. Reescalada	Numérica
EdadHoy	Edad actual. Reescalada	Numérica
IntervaloEdadHoy	["18-34", "35-64", "65+"]	Categórica
IntervaloEdadIn	["18-34", "35-64", "65+"]	Categórica
MesesAdentro	Reescalada	Numérica
incumplhist	Incumplimiento histórico	Numérica
AportePrevio	Monto de aporte justo antes del aumento. Reescalada	Numérica
porc_ult_aum	Porcentaje de aumento	Numérica
Binaria	1 si hay aumento	Binaria
Confort	1 si participó en la campaña	Binaria
SOAP18	1 si participó en la campaña	Binaria
SOAP19	1 si participó en la campaña	Binaria
participa	[0, 3] Cantidad campañas en que participa. Reescalada	Entera
RelProv	AportePrevio/PromedioProvincia. Reescalada	Numérica

Fuente: Elaboración propia

7.3. Análisis descriptivo

7.3.1. De los rangos altos de aporte

En esta sección se presenta un análisis descriptivo de los amigos pertenecientes a los segundo y tercer rangos de aporte, de $[30.001,\ 100.000]$ y $[100.001, 500.000]$.

En el Gráfico 5 se puede ver cómo están distribuídos los montos de las personas en el segundo rango. Se puede ver que la mayor parte están hasta los $50.000 y luego una cantidad considerable en el resto del rango,

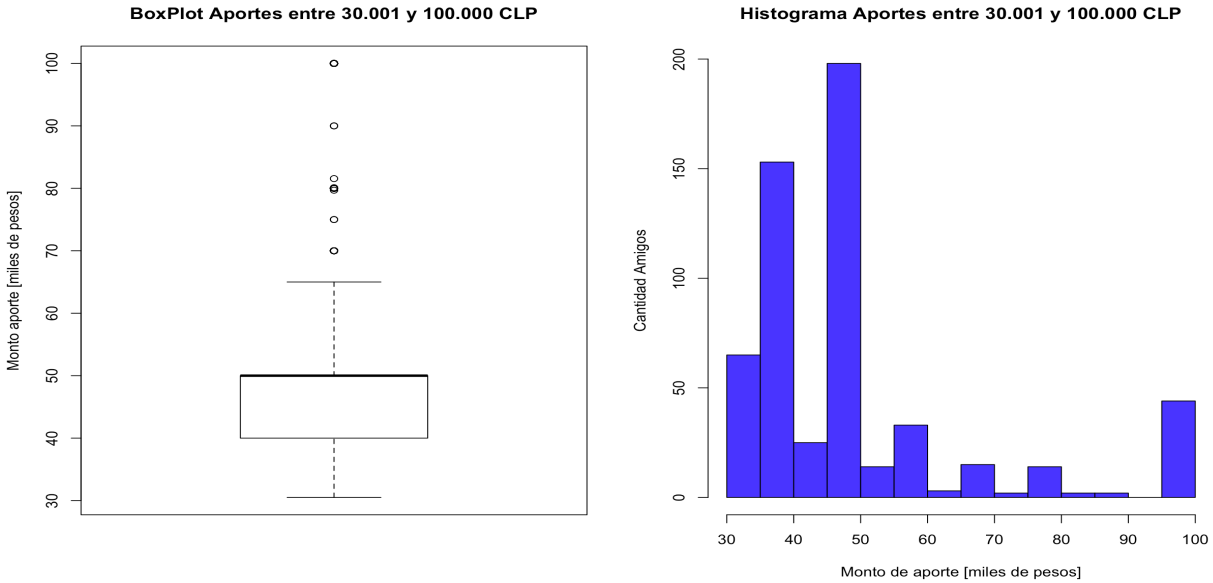
aumentando la concentración de personas en el valor $100.000, siendo 44 personas con este valor suscrito. Las cantidades de socios del segundo y tercer rango son 570 y 32 respectivamente.

A diferencia del segundo rango, en el tercero, el rango $[100.001, 500.000]$, en el Gráfico 6, posee todas las personas concentradas en la primera parte del rango, hasta los $300.000, teniendo 3 personas en este último valor[7], luego solo una persona con un monto de aporte de $500.000, por lo que este puede considerarse un *outlier*[8].

Si se observa el Gráfico 7 y el Gráfico 8 se pueden ver las cantidades de amigos de los dos rangos altos de aporte, separados por edad y por sexo. Una de las primeras observaciones que se aprecian es la marcada presencia de socios masculinos, representando el 63,6% en el segundo rango y el 78,1% en el tercero. Este comportamiento, como se verá más adelante, no es el mismo para el primer rango de aporte.

Por otra parte, se puede observar que no existen socios en estos rangos de bajas edades, teniendo 28 años el menor socio en el segundo rango y 44 años en el tercero. Esto se puede explicar por el aumento del poder adquisito que se presenta comunmente en esas edades.

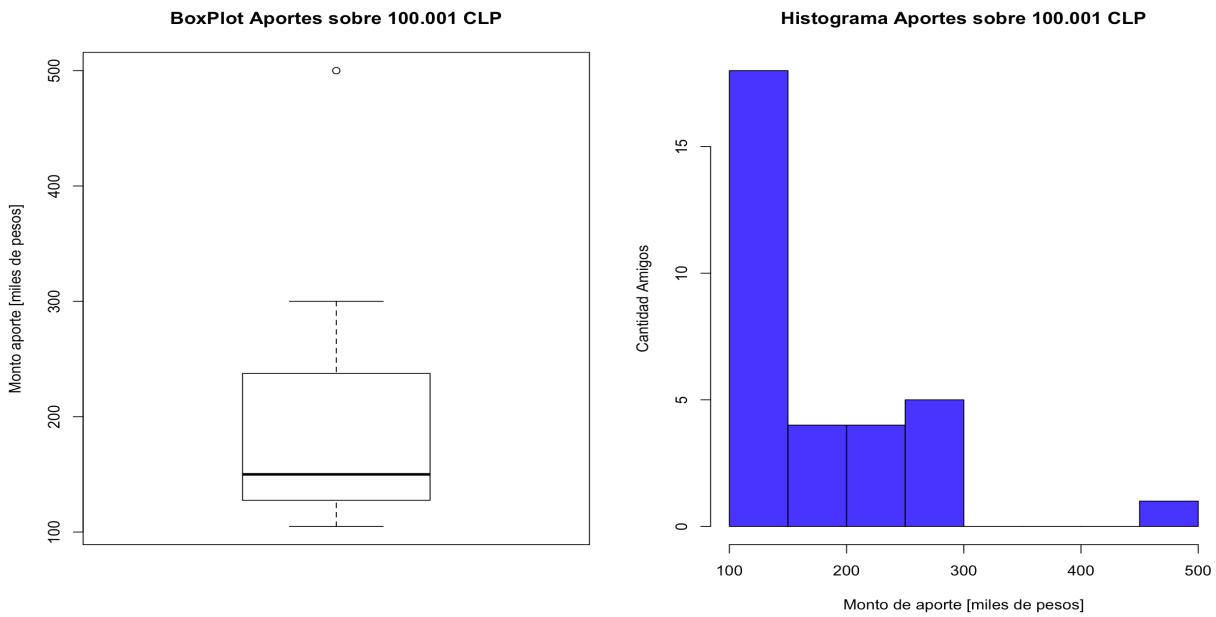
Gráfico 5 - Boxplot e histograma. Segundo rango de aporte



Fuente: Elaboración propia

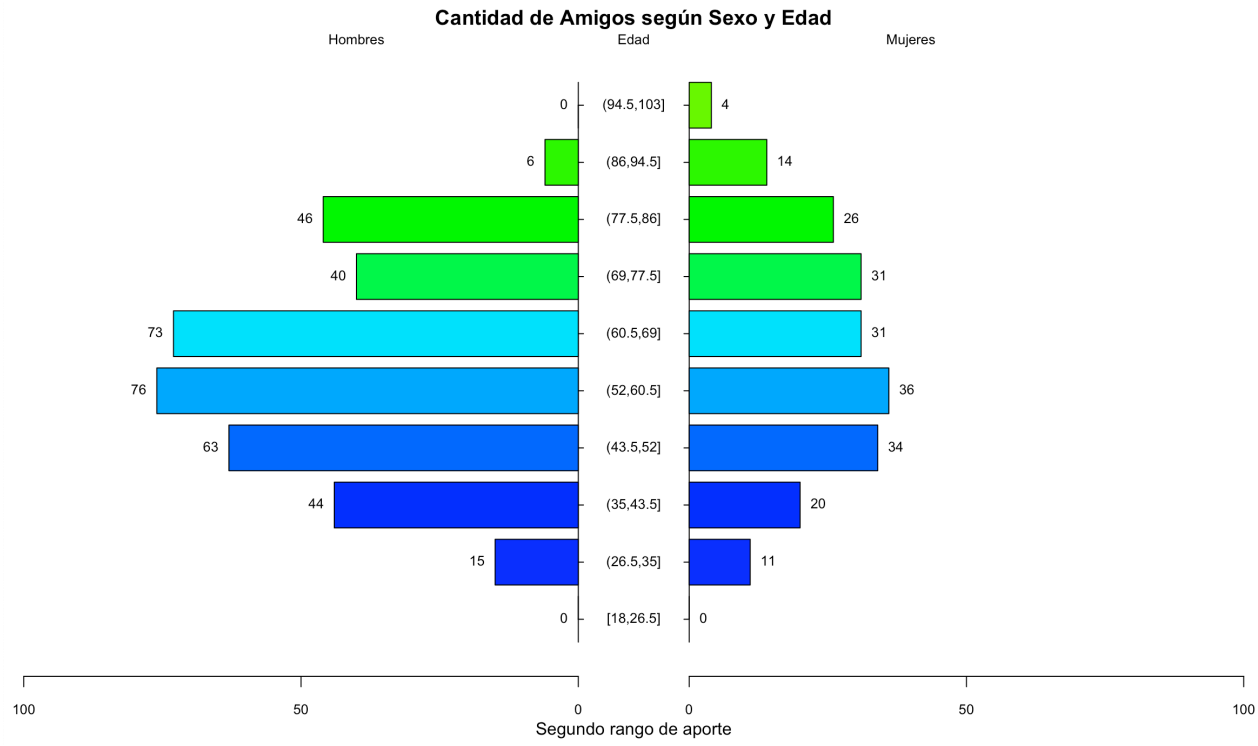
[7] En el Gráfico 6, la cuarta barra considera a las personas entre $250.001 y $300.000, ambos valores inclusive.
[8] Valor atípico

Gráfico 6 - Boxplot e histograma. Tercer rango de aporte



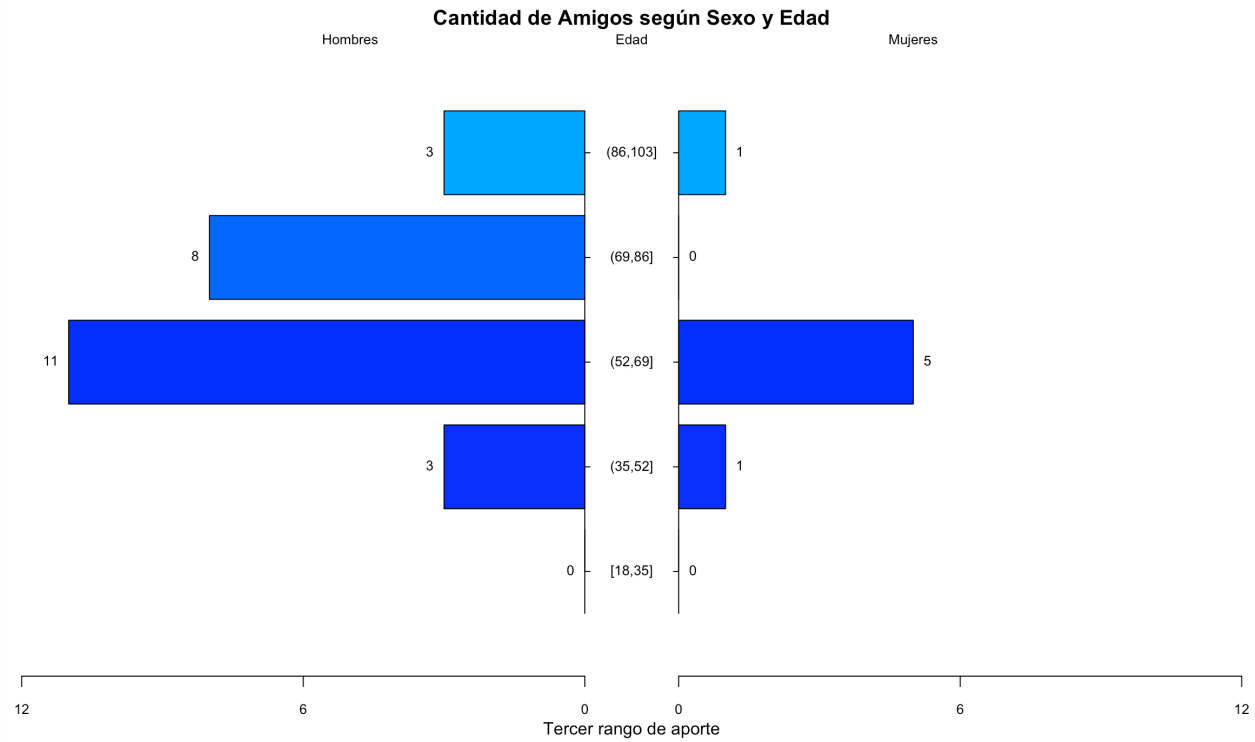
Fuente: Elaboración propia

Gráfico 7 - Cantidad de Amigos según Edad y Sexo. Segundo rango de aporte.



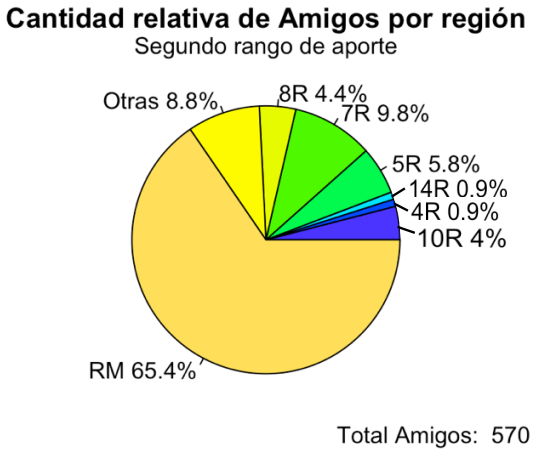
Fuente: Elaboración propia

Gráfico 8 - Cantidad de Amigos según Edad y Sexo. Tercer rango de aporte



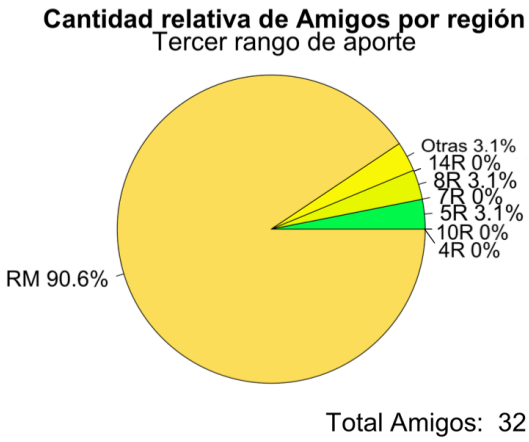
Fuente: Elaboración propia

Gráfico 9 - Amigos por región. Segundo rango de aporte



Fuente: Elaboración propia

Gráfico 10 - Amigos por región. Tercer rango de aporte



Fuente: Elaboración propia

Observando los Gráfico 9 y Gráfico 10 se pueden observar las cantidades porcentuales de los socios por región. La más evidente observación es la gran concentración de socios en la Región Metropolitana, siendo casi la totalidad de esa región en el tercer rango de aportes. Como se mostrará en el siguiente

apartado, en el primer rango de aportes se presenta este efecto, pero en un menor grado.

Con respecto a la cantidad de aumentos observados en la ventana de estudio, se tiene que, para el segundo rando de aportes, el 23% de socios tuvieron un aumento (133 de 570), el monto de aumento fue por un promedio de $18.192, con una desviación estándar de $14.674 y una correlación con *AportePrevio* de -0,38.

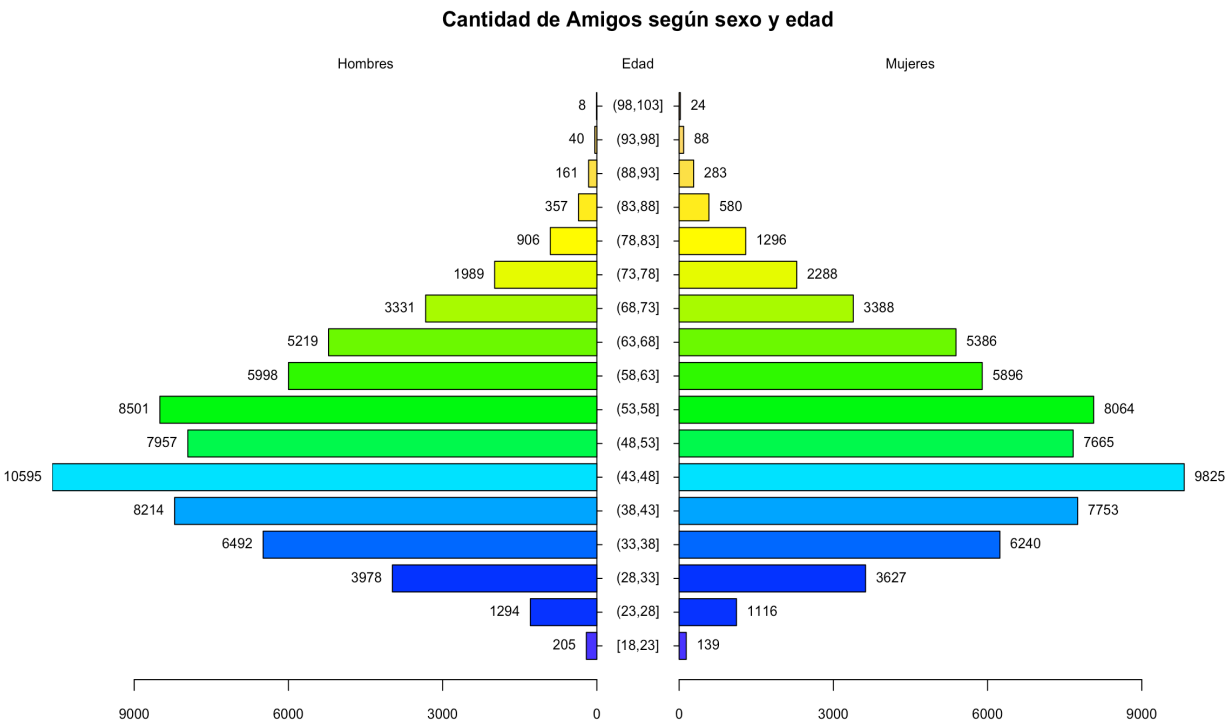
En el tercer rango de aporte se observó que el 25% de los socios(8 de 32) hicieron un aumento de aporte en la ventana de estudio. Estos aumentos fueron por un promedio de $42.875 con una desviación estándar de $37.517 y una correlación con *AportePrevio* de -0,36.

Al respecto de los aumentos, se puede observar que en ambos rangos el comportamiento es relativamente similar, en cuanto al porcentaje de personas que hicieron aumentos, la gran dispersión de estos aumentos que se observa al comparar el promedio con la desviación estándar. Además que se puede inferir que a menor *AportePrevio* mayor es el monto del aumento.

7.3.2. Del rango de aporte [0, 30.000]

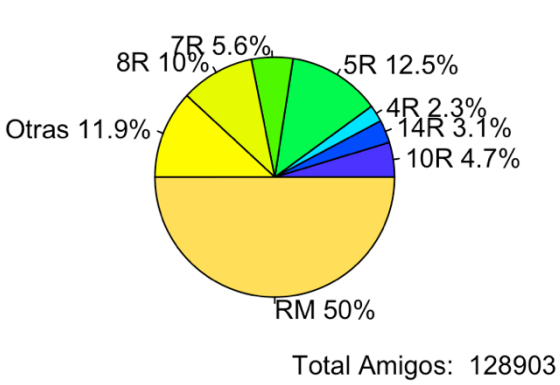
En el Gráfico 11 se presenta la cantidad de Amigos a nivel nacional, separados hombres y mujeres y distribuídos por rango de edad. Una primera observación es que la cantidad de hombres y mujeres es muy similar en todos los rangos etáreos, siendo las diferencias de cada rango inferior al 1% de la población, así, para efectos estadísticos se pueden considerar iguales. Un punto que llama la atención es la gran cantidad de personas entre los rangos (38, 58] que concentra el 53,2% de la cantidad de amigos. Cabe destacar que esta distribución se mantiene en todas las regiones (ver Gráfico 29 en anexos, página 66)

Gráfico 11 - Cantidad de Amigos según sexo y edad



Fuente: Elaboración propia

Gráfico 12 - Cantidad relativa de Amigos por región



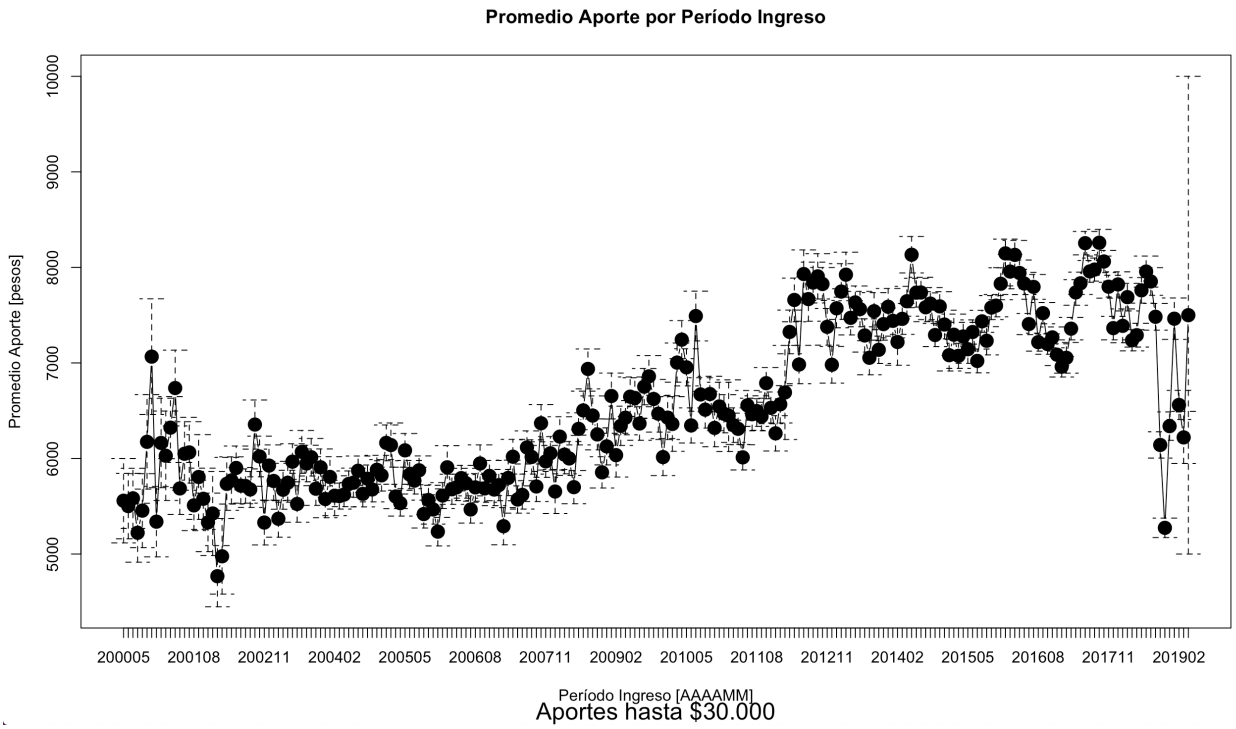
Fuente: Elaboración propia

En el Gráfico 12 se ve que la Región Metropolitana concentra el 50% de los Amigos de FLR. Se puede ver, además, que la cuarta, décima y catorceava región tienen participaciones de menos del 5%, además la séptima sobrepasa marginalmemte este valor. Luego están la quinta y octava región con el 12,5% y 10% respectivamente, cantidades que son relativamente considerables. Finalmente, hay un 11,9% pertenecientes a otras regiones de Chile, regiones en que FLR no tienen presencia, por lo que no se hace diferencia entre estas regiones.

En este rango de aporte existen 15 mil personas que realizaron aumentos, lo que representa un 11,7% del total. Estos aumentos fueron por un promedio

de $2.996 y tienen una desviación estándar de $2.469 y una correlación con *AportePrevio* de 0,32. Cabe destacar que, a diferencia de los rangos altos de aporte, en este caso se tiene una correlación positiva con el aporte previo, lo que indica que a mayor aporte previo sería mayor el monto del aumento. Además, el porcentaje de personas que realizan aumentos, en este rango, es menor que la mitad del porcentaje de los rangos anteriores.

Gráfico 13 - Promedio de Aporte por Período Ingreso



Fuente: Elaboración propia

El Gráfico 13 muestra cómo ha ido cambiando el promedio de aporte en el tiempo, es decir, el promedio de aporte de todos los amigos que ingresaron en un período. Cada uno de los valores promedios obtenidos entre los períodos 2000-05 (mayo del 2000) y 2019-03 (marzo del 2019), fueron obtenidos con entre 100 y 1000 observaciones, por lo que, se puede considerar que hay una leve tendencia a la alta.

Analizando la variable que indica el porcentaje de aumento, en la Tabla 8 se pueden ver los percentiles del 88 al 100. De esto se puede comprobar que al menos un 11% de la población ha hecho un aumento en la ventana de tiempo que se está analizando. No obstante, lo que llama la atención, es el salto del percentil 99 al 100, que pasa de un aumento del 100% a un aumento del 950%, lo que hace pensar que este último valor podría ser un *outlier*.

Sin embargo, en el Gráfico 14 se muestra la distribución acumulada del porcentaje de aumento, la linea roja indica un aumento del 100%. Se puede

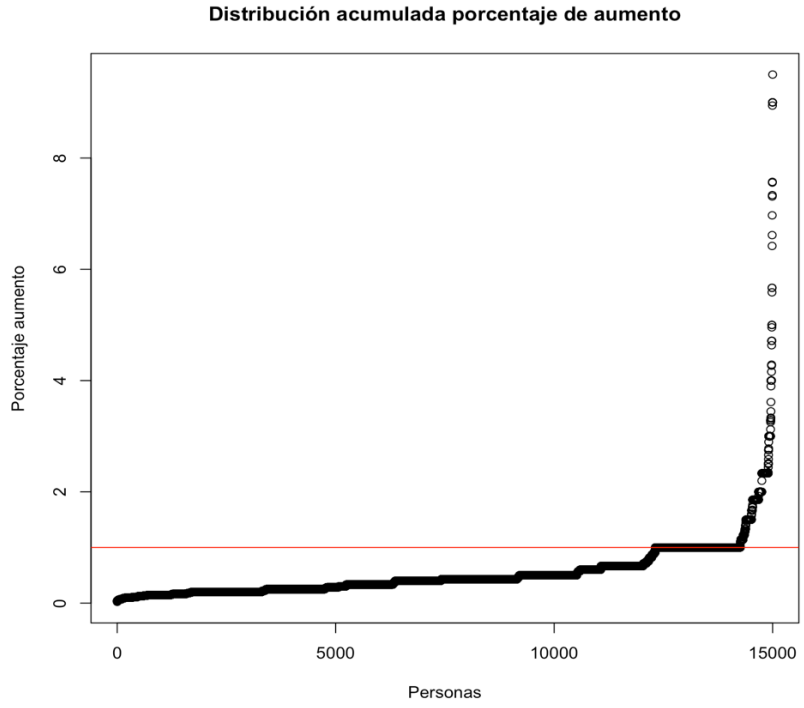
ver que hay muchos casos con aumentos sobre este valor, además que el crecimiento es paulatino, por lo que se puede descartar que estos valores sean *outliers*.

Tabla 8 - Percentiles [88-100] porcentaje de aumento

88%	89%	90%	91%	92%	93%	94%
0.000	0.142	0.200	0.250	0.272	0.333	0.400
95%	96%	97%	98%	99%	100%	
0.428	0.500	0.666	1.000	1.000	9.500	

Fuente: Elaboración propia

Gráfico 14 - Distribución acumulada porcentaje aumento



Fuente: Elaboración propia

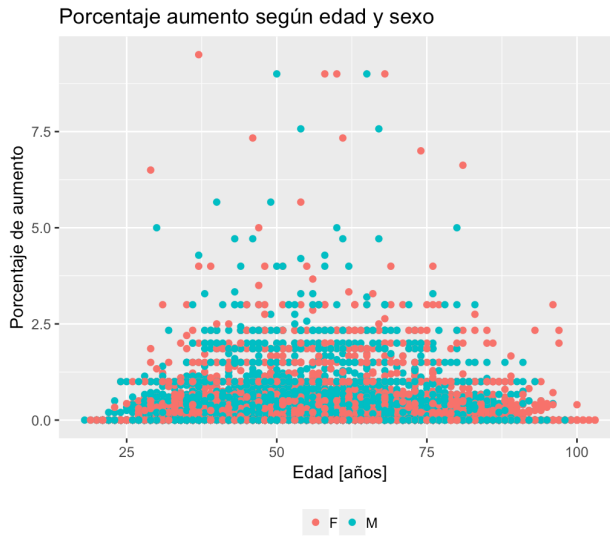
7.3.3. Análisis Bivariado

La variable porcentaje de aumeto es una de las que se busca incluir como dependiente en los modelos, es la que se quiere entender el comportamiento, por lo que se hará un análisis bivariado que incluya esta variable dependiente con otras variables que puedan resultar relevantes.

Así, en el Gráfico 15 se puede ver cómo se comporta el porcentaje aumento según la edad actual del socio y su sexo. Se puede ver que la variable se

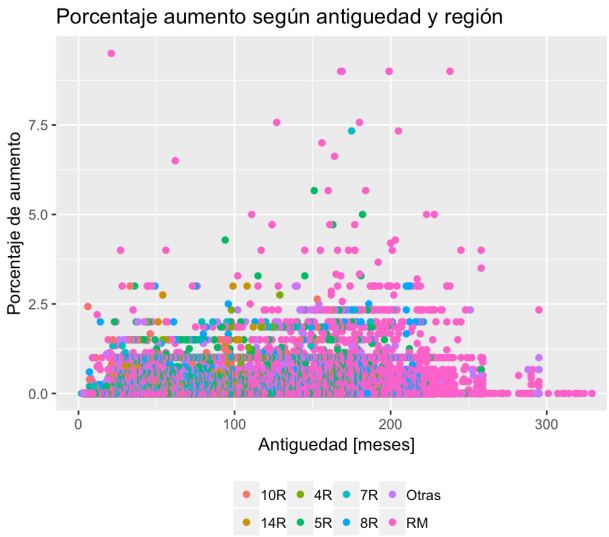
distribuye homogeneamente tanto en la edad como en el sexo. Por otra parte, el Gráfico 16 muestra la variable distribuída según la antigüedad en el programa y según la región del socio, igualmente que en el caso anterior, se puede ver que la variable se distribuye homogéneamente según la antigüedad.

Gráfico 15 - Porcentaje aumento según edad y sexo



Fuente: Elaboración propia

Gráfico 16 - Porcentaje de aumento según antiguedad y región

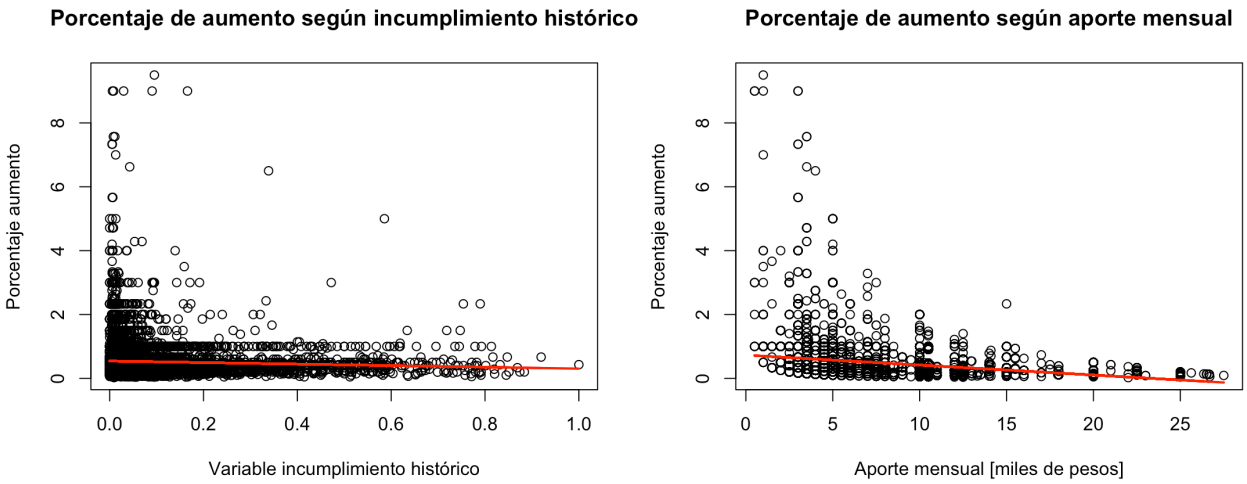


Fuente: Elaboración propia

Continuando con el análisis bivariado, en el lado izquierdo del Gráfico 17 se puede ver cómo se comporta el porcentaje de aumento según la variable incumplimiento histórico. De ese gráfico se puede ver que para valores pequeños de aumento, esto es, hasta 100% de aumento, el incumplimiento histórico se distribuye homogéneo. No obstante, y acorde a la intuición, los aumentos de mayor porcentaje se concentran en valores bajos de incumplimiento. Sin embargo, no se ve una tendencia clara.

Analizando el gráfico de la derecha del Gráfico 17 se ve que el porcentaje de aumento, para valores sobre 200%, se concentra en las donaciones hasta $10.000. La línea roja en este gráfico representa una regresión lineal que relaciona ambas variables, al tener pendiente negativa se puede deducir que están correlacionadas de manera inversa, esto es, mientras menor aporte mensual, mayor es el porcentaje de aumento, no obstante, no es una tendencia fuerte, ya que la recta está muy cerca de ser horizontal.

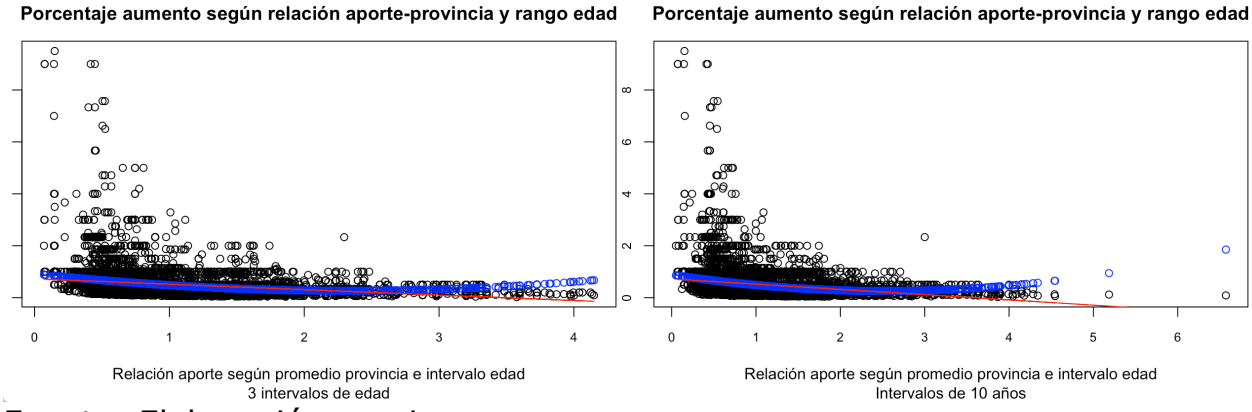
Gráfico 17 - Porcentaje aumento según incumplimiento histórico y aporte mensual



Fuente: Elaboración propia

Luego, analizando el porcentaje de aumento con la relación de *AportePrevio* frente a un promedio de aporte según edad y provincia, se puede ver, en el Gráfico 18, que los grandes porcentajes de aumentos se dan, mayoritariamente, en el intervalo [0,1) de relación frente al promedio, esto es, las personas que están bajo el promedio. Adicionalmente, se ve que al ajustarse la curva a un comportamiento cuadrático, existe una pequeña curva, que tiende a pronunciarse un poco más en el gráfico de la derecha, esto lleva a pensar que la relación entre estas dos variables puede no ser lineal.

Gráfico 18 - Porcentaje de aumeto según relación aporte provincia y rango edad



Fuente: Elaboración propia

7.3.4. De las campañas

En la campaña del SOAP en marzo de 2018 se vendieron 1.865 pólizas a personas inscritas al programa. En marzo de 2019 se vendieron 1.562, de los

cuales 424 habían comprado el año anterior. Por lo anterior, un 22,73% de los que compraron póliza el 2018, repitieron la compra el 2019, además, aunque considerando la base total, en ambos años un 1,2% de las personas participaron, las ventas de pólizas bajaron 16,24% de un año a otro.

En agosto de 2018, en la campaña "7 razones", monto que se comenzó a devolver ese mes. En esta campaña participaron 2.981 personas, esto es, 2,3% de la población. De ellos, 43 y 44 compraron SOAP el año 2018 y 2019 respectivamente.

7.4. Minería de Datos

Se presentan en esta sección 3 modelos realizados sobre la base. Cabe destacar que los modelos Logit y Random Forest, como se comentó en el apartado 4.1, buscan una predicción, que en este caso es una estimación puntual, por lo que se encuentran en un nivel de investigación del tipo predictivo. Es por esto que previo a la construcción de estos modelos, se extrajo el 30% de la data aleatoriamente sin reemplazo, para dejarla como base para testeo. Esta base nunca fue utilizada en la construcción del modelo, por lo que tampoco pasó por los balanceo que se realizarán más adelante.

7.4.1. Modelo de Regresión Lineal

Para la creación de regresiones lineales existen distintas estrategias, lo más utilizado es ir agreganto o quitando regresores para analizar cuáles son los cambios en los coeficientes de las otras variables y con el R^2, así ir sacando conclusiones hasta el modelo final. Ya que el objetivo final de esta memoria no es llegar a un modelo regresivo, sino que, el diseño de experimentos, es que se ira concluyéndo de varios modelos para alimentar el diseño experimental.

El primer modelo (ver Modelo 1) fue una regresión lineal con las principales variables explicativas que se manejan. Lo primero que llama la atención es el bajo valor del coeficiente R^2, este valor indica que el modelo está explicando el 2% de la varianza, lo que no es representativo, por lo que utilizar este modelo para hacer predicciones sería erróneo.

Modelo 1 - Regresión Lineal

Fórmula = porc_ult_aum ~Genero + AportePrevio + mesesadentroprevio + edadhoy + incumplhist
Data = basegigante30k
Coeficientes=

	Estimado	Error estándar	Significancia
(Intercepto)	0.028	0.002	***
GeneroM	0.007	0.001	***
AportePrevio	-0.044	0.005	***
mesesadentroprevio	0.123	0.003	***
edadhoy	0.033	0.004	***
incumplhist	-0.072	0.003	***

Códigos de significancia
0 *** 0.001 ** 0.01 * 0.05 · 0.1 1
Multiple R^2 = 0.024 R^2 Ajustado = 0.024

Sin embargo, se ve que todas las variables son significativas, de aquí se puede desprender que el porcentaje de aumento sigue una tendencia que está relacionada con las variables observables. Además, al estar todas las variables numéricas reescaladas entre 0 y 1, son comparables sus coeficientes. Así, se puede decir que lo que más influye en el porcentaje de aumento es la variable de antigüedad de manera positiva, luego incumplimiento de manera negativa, aporte previo de manera negativa y edad de manera positiva. Además, se puede desprender que los hombres tienen mayor porcentaje de aumento.

7.4.2. Modelo Logit

Al no llegar a un modelo con suficiente significancia global, se comenzó a utilizar la variable dependiente binaria, esto ya que predecir si una persona acepta o no un aumento es más fácil de estimar en comparación a estimar cuánto va a aumentar. Así, se obtuvo el segundo modelo (ver Modelo 2), un modelo Logit, con las mismas variables dependientes que el modelo anterior.

Modelo 2 - Logit

Fórmula = binaria ~Genero + AportePrevio + mesesadentroprevio + edadhoy + incumplhist
Data = basegigante30k.entreno
Coeficientes=

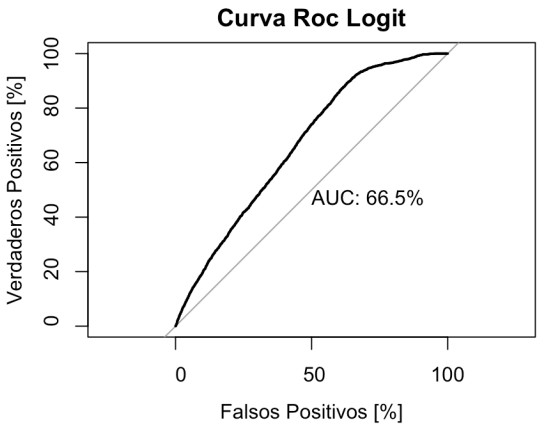
	Estimado	Error estándar	Significancia
(Intercepto)	-2.533	0.039	***
GeneroM	0.067	0.019	***
AportePrevio	0.584	0.073	***
mesesadentroprevio	1.134	0.058	***
edadhoy	0.752	0.066	***
incumplhist	-3.751	0.110	***

Códigos de significancia
0 *** 0.001 ** 0.01 * 0.05 · 0.1 1
AIC = 70021

En el Gráfico 19 se presenta la curva ROC obtenida desde la predicción de este modelo en la base de entrenamiento. Lo primero que se aprecia es que tiene un valor de $AUC = 66.5\%$, esto hace que la curva esté por sobre la diagonal que marca la elección al azar, pero sigue siendo bajo para considerar que tiene algún poder de predicción.

No obstante, al igual que en el modelo anterior, con los valores de significancia se puede ver la tendencia que sigue la variable dependiente según los regresores, además de la fuerza con la que va cambiando. Los coeficientes de los regresores son un simil de la pendiente, por lo que mientras más alejado del 0 esté el valor, más rápido va a aumentar o disminuir la variable dependiente.

Gráfico 19 - Curva ROC Logit



Fuente: Elaboración propia

Se puede desprender de los coeficientes que la variable que más influye en que los socios acepten o no un aumento es, de manera negativa, el incumplimiento, luego la antigüedad de manera positiva, la edad de manera positiva y el aporte de manera positiva. Igual que en el modelo anterior, también se puede desprender que los hombres tienen una mayor tendencia a aceptar aumentos que las mujeres.

7.4.3. Variable AportePrevio

Llama la atención el cambio de signo de la variable *AportePrevio* entre un modelo y otro, ya que en el modelo de regresión numérica afecta de manera negativa y para la regresión logística de manera positiva. En este punto es importante recordar la manera en la que se calculó la variable dependiente de la regresión numérica, *porc_ult_aum*, para las personas que sí presentaban algún aumento de aporte se calculó como:

$$P_{ultaum} = \frac{AporteActual - AportePrevio}{AportePrevio}$$

Al estar dividiendo por *AportePrevio* se normaliza el monto del aumento (por ejemplo $5.000) con respecto al monto incial de dónde se está aumentando, así, se captura la diferencia entre dos personas que aumentan el mismo monto, pero que aportan inicialmente distinto (por ejemplo no es lo mismo subir de $5.000 a $10.000, aumento el doble, que de $25.000 a $30.000, que aumenta el 20%). Pero además, este efecto provoca que, matemáticamente, la variable dependiente esté correlacionada de manera negativa con la variable *AportePesos*, eso explica el signo negativo en la regresión lineal.

Al comparar con el Logit, en el cual, la variable dependiente no utilizó la variable *AportePrevio*, se captura este efecto de manera más sincera. Aun así, resulta interesante agregar la variable $AportePrevio^2$ como una manera de capturar este efecto, así, se presentan una nueva regresión lineal (ver Modelo 3) y un nuevo logit (ver Modelo 4).

Se puede ver que, en ambas regresiones, el coeficiente de la variable *AportePrevio* es positivo y el coeficiente de la variable cuadrática es negativo, esto con niveles de significancia de al menos 99%. Además, se puede ver que el módulo del coeficiente de la variable cuadrática es mayor en ambos modelos, lo que implica que tiene una mayor incidencia en las variables dependientes. Esto explica el cambio de signo en el coeficiente entre los dos primero modelos, se puede deducir, como se suponía en el análisis descriptivo, que la variable AportePrevio tiene un componente cuadrático que afecta la variable dependiente en ambos casos. Traduciendo de lo matemático a lo real, se interpreta que, para niveles bajos[9] de la variable, a mayor aporte, mayor es la tendencia a aceptar un aumento y mayor es el porcentaje de aumento, pero llega un punto, un nivel de aporte, en que, para valores más altos, se comportan de manera negativa. Es decir, hay un punto de inflexión en la variable AportePrevio, que hace que seguir aumentando, disminuya la tendencia hacia el aumento.

Modelo 3 - Regresión lineal AportePesos2

Fórmula = binaria ~Genero + AportePrevio + AportePrevio2 + mesesadentroprevio + edadhoy + incumplhist
Data = basegigante30k.entreno
Coeficientes=

	Estimado	Error estándar	Significancia
(Intercepto)	0.016	0.003	***
GeneroM	0.007	0.001	***
AportePrevio	0.036	0.014	*
AportePrevio2	-0.110	0.018	***
mesesadentroprevio	0.128	0.003	***
edadhoy	0.033	0.004	***
incumplhist	-0.072	0.003	***

Códigos de significancia
0 *** 0.001 ** 0.01 * 0.05 · 0.1 1
Multiple R^2 = 0.024 R^2 Ajustado= 0.024

Modelo 4 - Regresión Logística AportePesos^2

Fórmula = binaria ~Genero + AportePrevio + AportePrevio2 + mesesadentroprevio + edadhoy + incumplhist
Data = basegigante30k.entreno
Coeficientes=

	Estimado	Error estándar	Significancia
(Intercepto)	-3.282	0.052	***
GeneroM	0.055	0.019	**
AportePrevio	6.025	0.273	***
AportePrevio2	-7.984	0.423	***
mesesadentroprevio	1.429	0.060	***
edadhoy	0.781	0.066	***
incumplhist	-3.869	0.113	***

Códigos de significancia
0 *** 0.001 ** 0.01 * 0.05 · 0.1 1
AIC = 69550

Cabe destacar que la significancia global de los modelos 3 y 4 es prácticamente la misma que los modelos 1 y 2, esto basado en los valores de R^2 para las regresiones lineales y el AUC (=66.7%, ver Gráfico 30 en 68) para las regresiones logísticas.

[9] Dependiendo de la transformación que se hizo para dejar entre 0 y 1

7.4.4. Clustering – Kmeans

Se comenzó analizando las correlaciones de las variables para elegir candidatos a formar clusters, en la Tabla 9 se muestran estas correlaciones. Los primero que se puede extraer, son las altas correlaciones entre ciertos grupos de variables, esto se debe a que contienen el mismo comportamiento. Así, existen grupos variables que tienen información sobre: la edad (*edadin* y *edadhoy*), sobre antigüedad (*MesesAporte* y *mesesadentro*) y sobre aporte (*AportePrevio*, *RelProvinciaTodos* y *RelProvincia3int*).

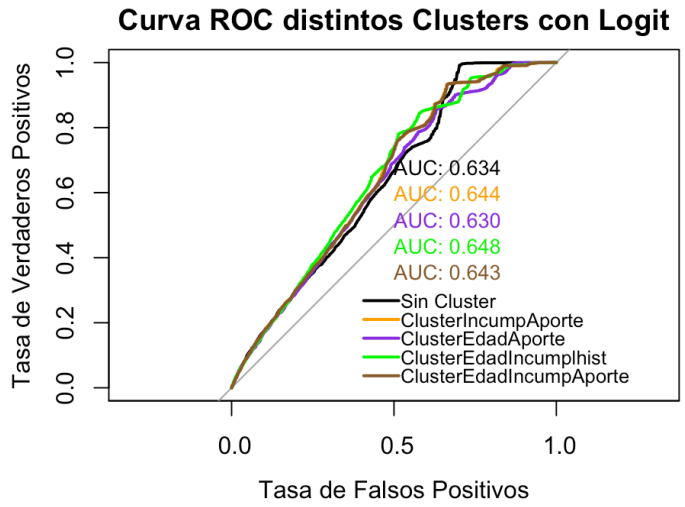
Tabla 9 - Correlaciones

	porc_ult_aum	edadin	edadhoy	MesesAporte	MesesAdentr	incumplhist	AportePrevio	RelProvTodos	RelProv3int
porc_ult_aum	1.000	0.031	0.070	0.116	0.102	-0.112	-0.049	-0.047	-0.051
edadin	0.031	1.000	0.915	-0.050	-0.071	-0.071	0.064	0.068	0.007
edadhoy	0.070	0.915	1.000	0.339	0.333	-0.224	-0.020	-0.014	-0.082
MesesAporte	0.116	-0.050	0.339	1.000	0.965	-0.556	-0.207	-0.202	-0.227
MesesAdentr	0.102	-0.071	0.333	0.965	1.000	-0.392	-0.206	-0.200	-0.225
incumplhist	-0.112	-0.071	-0.224	-0.556	-0.392	1.000	0.079	0.073	0.089
AportePrevio	-0.049	0.064	-0.020	-0.207	-0.206	0.079	1.000	0.990	0.984
RelProvTodos	-0.047	0.068	-0.014	-0.202	-0.200	0.073	0.990	1.000	0.994
RelProv3int	-0.051	0.007	-0.082	-0.227	-0.225	0.089	0.984	0.994	1.000

Fuente: Elaboración Propia

Lo ideal para formar clusters es encontrar variables con correlación baja, al destacar las correlaciones que en módulo son menores a 0,1, se ve que las variables del antigüedad están por sobre este valor con las otras (salvo *edadin*), con esto se puede pensar que no son buenas candidatas a fomar clusters.

Gráfico 20 - Curva ROC Logit Clusters



Fuente: Elaboración Propia

De las restantes, se destacaron variables de edad en azul, de incumplimiento en amarillo y de aporte en rojo. Las correlaciones entre ellas se destacaron según la combinación de los respectivos colores[10]. Para hacer más fácil la lectura, se referirá a cada cluster por estos colores.

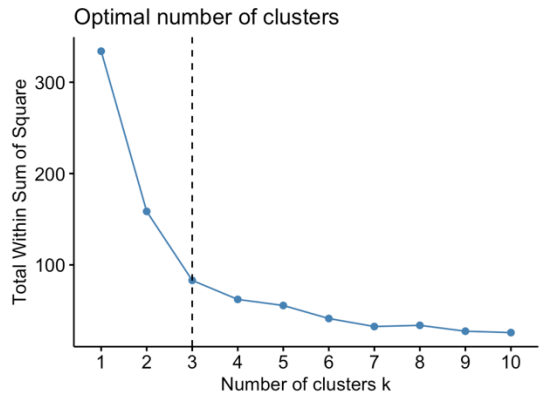
La selección de un cluster se basa principalmente en la utilidad que este presenta, además de cómo aporta al modelo. A continuación se

[10] Rojo + Amarillo = Naranjo, Rojo + Azul = Morado, Azul + Amarillo = Verde y Todos = Café

presenta un gráfico con la curva ROC de 5 modelos, todos con las variables *Genero* y *mesesadentroprevio* a modo de control, más, una variable que indica el cluster respectivo.

Lo primero que se puede apreciar es la similitud entre los distintos valores de AUC. Al igual que en el caso anterior, no permite afirmar que tiene un poder predictivo, pero sí permite apreciar una tendencia. Por ejemplo, se puede descartar el cluster morado ya que baja el AUC.

Gráfico 21 - Cantidad de Clusters



Fuente: Elaboración propia

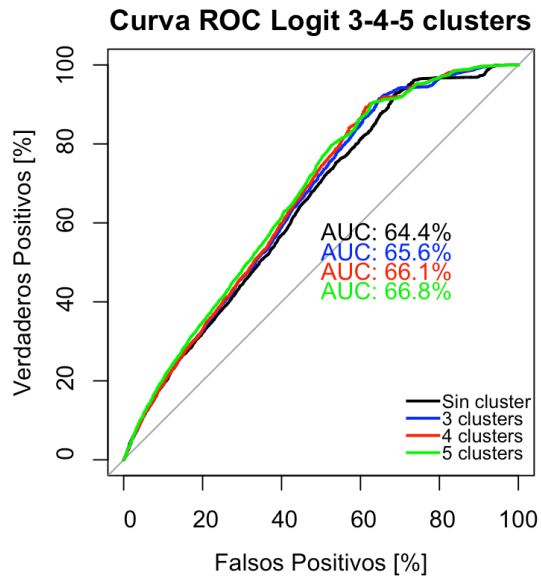
Al comparar el cluster naranjo con el café, la diferencia es que este último tiene incorporada una variable de la edad. Al agregarle esta variable baja el valor del AUC, por lo que se puede descartar este cluster también. Finalmente los clusters tentativos son el verde y el naranjo, con variables *incumplhist* y de *Aporte* para el primero e *incumplhist* y de *edad* para el segundo.

Ya que las variables de edad son muy intuitivas para llevar a la realidad, además que entregan mucha información si se trabajan con intervalos de edad, se las prefirió dejar fuera del cluster, por lo que, en adelante, se trabajarán con el cluster naranjo.

En el **¡Error! No se encuentra el origen de la referencia.** se puede ver cómo va cambiando la suma de los cuadrados de las distancias de a los centroides del cluster. Se puede ver que, al pasar de 3 a 4, la distancia que se reduce es mínima, por lo que, por el criterio del codo, la cantidad óptima de clusters es 3.

Por otra parte, si se observa el Gráfico 22, se puede ver los distintos valores de AUC para las distintas cantidades de clusters. Se puede ver que son muy similares entre ellas, adicionalmente, se pueden ver en la Tabla 10, gráficos

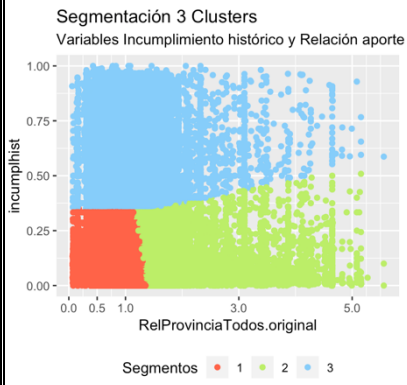
Gráfico 22 - Curva ROC Logit 3-4-5 clusters



Fuente: Elaboración propia

de los segmentos para las distintas cantidades de clusters. Para una mejor interpertación, en estos gráficos se utilizó la variable original y no su versión escalada, por esa razón, el eje x no está entre 0 y 1.

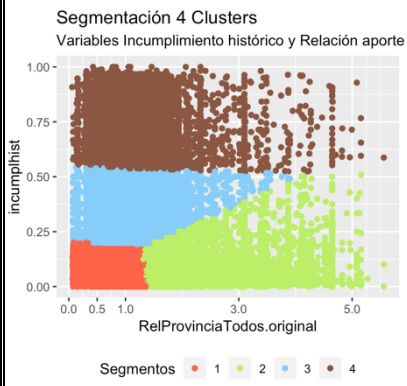
Tabla 10 - 3-4-5 Clusters



Fórmula = binaria ~ Genero + intervaloedadhoy + mesesadentroprevio + AportePrevio + clust3IncumpRelProvT
Data = basegigante30k.entreno
Coeficientes=

	Estimado
(Intercepto)	-3.230
GeneroM	0.052
intervaloedadhoy35-64	0.563
intervaloedadhoy65+	0.751
mesesadentroprevio	1.698
AportePrevio	1.088
clust3IncumpRelProvT2	-0.235
clust3IncumpRelProvT3	-1.551

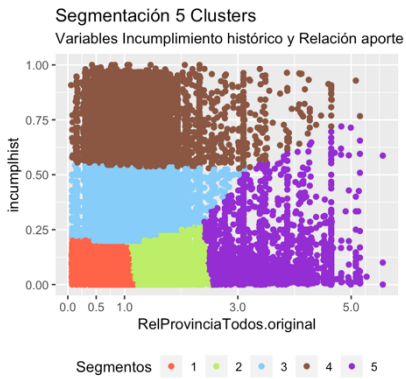
AIC = 71030



Fórmula = binaria ~ Genero + intervaloedadhoy + mesesadentroprevio + AportePrevio + clust4IncumpRelProvT
Data = basegigante30k.entreno
Coeficientes=

	Estimado
(Intercepto)	-3.037
GeneroM	0.052
intervaloedadhoy35-64	0.520
intervaloedadhoy65+	0.701
mesesadentroprevio	1.471
AportePrevio	0.991
clust4IncumpRelProvT2	-0.226
clust4IncumpRelProvT3	-1.123
clust4IncumpRelProvT4	-2.190

AIC = 70520



Fórmula = binaria ~ Genero + intervaloedadhoy + mesesadentroprevio + AportePrevio + clust5IncumpRelProvT
Data = basegigante30k.entreno
Coeficientes=

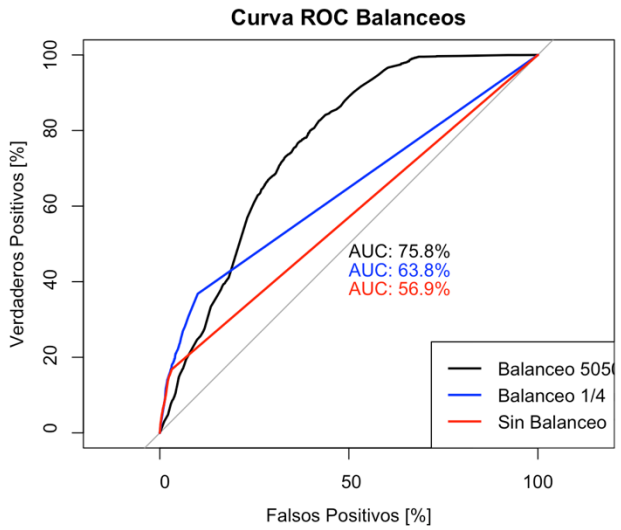
	Estimado
(Intercepto)	-3.222
GeneroM	0.047
intervaloedadhoy35-64	0.517
intervaloedadhoy65+	0.700
mesesadentroprevio	1.568
AportePrevio	0.945
clust5IncumpRelProvT2	-0.179
clust5IncumpRelProvT3	-1.157
clust5IncumpRelProvT4	-2.246
clust5IncumpRelProvT5	-1.407

AIC = 70390

Fuente: Elaboración propia

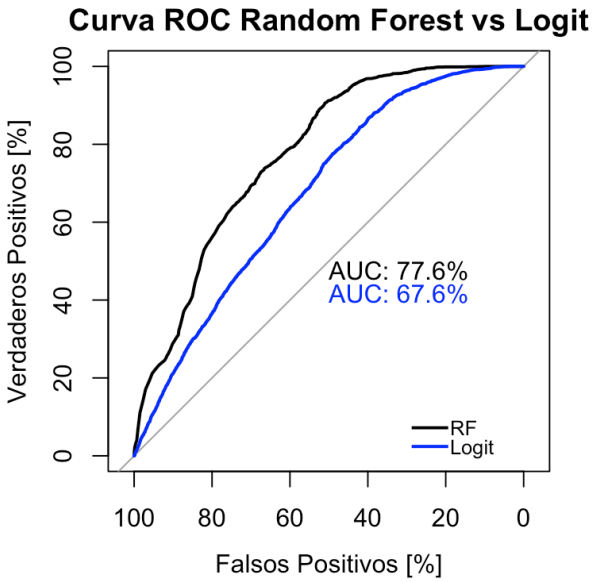
Se puede ver cómo se ordenan los clusters según su tendencia a aceptar o no el aumento de aporte. Según los coeficientes, en orden de mayor a menor propensión de aceptar aumento, los segmentos son: 1, 2, 3, 5 y 4, para los 5 clusters. El orden de los segmentos se mantiene para las otras cantidades. Es importante destacar el aumento en el coeficiente del segmento 4, que indica un mayor rechazo a aceptar un aumento.

Gráfico 23 - Curvas ROC distintos balanceos



Fuente: Elaboración propia

Gráfico 24 – Curva ROC RF vs Logit



Fuente: Elaboración propia

7.4.5. Modelo Random Forest

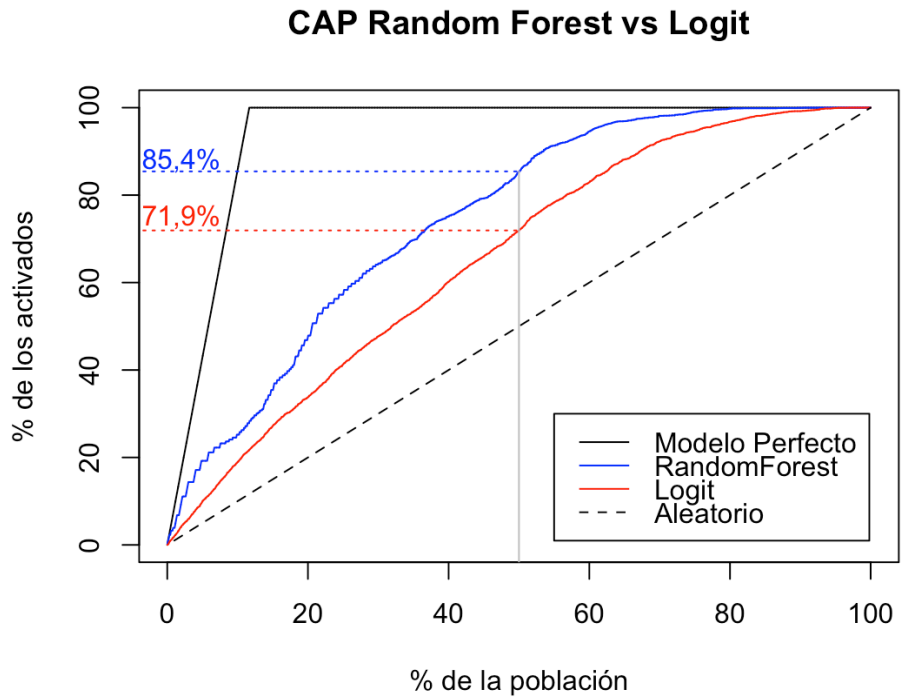
Como se explicó en el apartado 4.1.1.3-a, los modelos de RF son sensibles a las bases desbalanceadas, esto es, una concentración de una clase mucho mayor que otra. En este caso la proporción es, aproximadamente 1:8, es decir, 1 socio con aumento cada 8 socios sin aumento. Es por esto que, antes de analizar los modelos de RF, se procedió a correr un modelo genérico, para distintos balanceos de la base, las curvas ROC se presentan en el Gráfico 23.

En este gráfico se puede ver cómo afectan los distintos balanceos, primero en lo más evidente que es el valor del AUC, hace que sean mucho menos efectivo el uso de árboles. Además, luego de los primeros verdaderos positivos que logró identificar, posteriormente comienza a asignar de manera aleatoria, lo que explica lo recta que es el gráfico.

A continuación, se presenta el Gráfico 24 en el cual se hace una comparación, por medio de la curva ROC, de las predicciones obtenidas por los modelos Logit y RandomForest. Se puede ver que

el modelo Random Forest es el que mejor desempeño tiene por lo valores de AUC. Posteriormente, en el Gráfico 25, se confirma esto y se observa además, el porcentaje de activados que ambos modelos obtuvieron al 50% de los datos. El modelo de RF logró detectar un 85% de los activados al 50% de los datos, esto confirma su poder predictor.

Gráfico 25 - Curva CAP RF vs Logit



Fuente: Elaboración propia

7.5. Diseño Experimental

La elección de los comportamientos que se quieren estudiar depende netamente del objetivo del investigador o de la estrategia a la cuál el estudio hace efecto, además de la factibilidad de su implementación. Por esta razón, es que, idealmente, deben ser escogidos en conjunto con la institución.

En este caso FLR no participó en esa etapa, esto por la escasa cantidad de variables comportamentales de los socios, y poca experiencia de la institución en lo referido a marketing directo, así, se prefirió privilegiar la factibilidad del experimento.

Así, se definieron tres hipótesis a testear, de acuerdo a los modelos y análisis que se fueron desarrollando. La primera de estas nace a partir del modelo RandomForest, el cual logró un buen desempeño, por lo que la hipótesis es:

$$H_1: El\ modelo\ de\ RandomForest\ identifica\ personas\ con\ mejor\ tasa\ de\ respuesta$$

Dada la ventana de transacciones disponible, habían personas con más de un aumento. Así, surge la necesidad de utilizar una variable que tuviera los periodos que han pasado desde la última vez que se realizó un aumento de aporte. Desafortunadamente, con la información que existía, no se podía modelar tal variable ya que el número de personas con más de un aumento es muy bajo. No obstante, se cree que esta variable puede ser relevante en la tasa de respuesta, por lo que la segunda hipótesis es:

$$H_2: La\ variable\ PeriodosUltimoAumento\ influye\ en\ las\ tasas\ de\ respuesta$$

Finalmente, sería muy deseable poder llegar a una segmentación que fuese útil de alguna manera, esto por la fácil aplicación que puede tener este conocimiento. Así, la última hipótesis es:

$$H_3: Las\ tasas\ de\ respuesta\ de\ los\ clusters\ son\ distintas$$

7.5.1. H_1 → Modelo Random Forest

El modelo RandomForest que se desarrolló es del tipo clasificador, tiene como variable dependiente la variable *binaria* en la que se encuentra la información de si la persona hizo o no el aumento. Lo que se busca con este experimento es determinar si el modelo logra predecir de manera efectiva. Es por esto, que, formalmente, la hipótesis a testear es:

$$H_1: p_1 = p_2$$

Donde p_1 es la proporción de personas que aceptaron el aumento del grupo de tratamiento, mientras que p_2 es del grupo de control. Estos valores deben ser estimados ya que serán utilizados para el cálculo de los tamaños muestrales.

7.5.2. H_2 → PeriodosUltimoAumento

En este caso se quiere medir la influencia que tiene esta variable tanto en si una persona acepta hacer un aumento, como en la cantidad en que aumenta. Así, existe una hipótesis para proporciones (p) y otra para medias (μ). Además, se quiere trabajar con 3 posibles valores para los periodos previo al último aumento: 12, 18 y 24. Así, formalmente las hipótesis a testear son:

$$H_2^p: p_{12} = p_{18} = p_{24}$$
$$H_2^\mu: \mu_{12} = \mu_{18} = \mu_{24}$$

7.5.3. H_3 → Cluster

En este caso, se busca saber si existe alguna diferencia estadísticamente significativa en la respuesta de los distintos segmentos encontrados por k-means. Para crear las categorías a estudiar, se agruparon las categorías del cluster según su similitud en el coeficiente de la regresión logística expuestos en la Tabla 10. Así, se buscan diferencias en las proporciones y medias de los

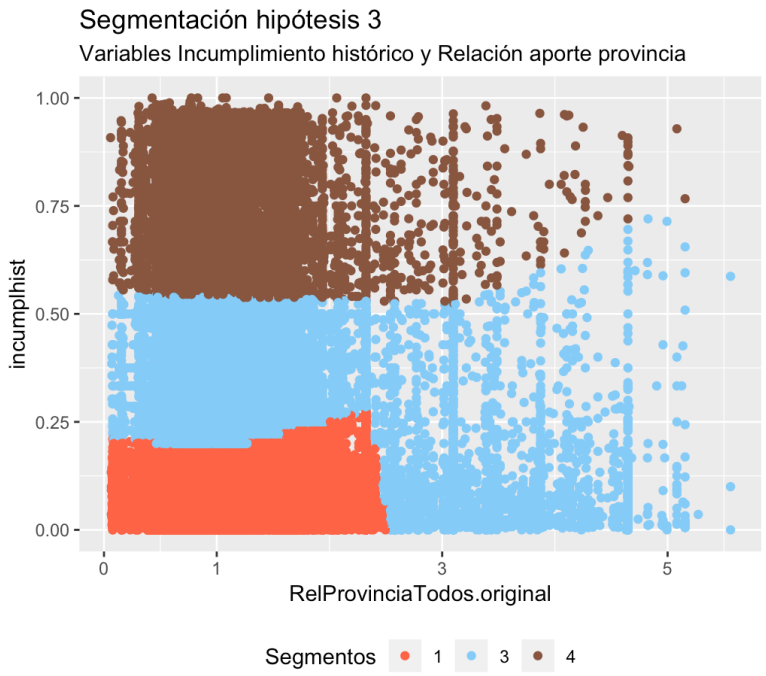
segmentos 1-2, 3-5 y 4, en adelante llamados 1, 3 y 4 respectivamente. La nueva segmentación se muestra en el Gráfico 26.

Así, formalmente, las hipótesis que se busca probar son:

$$H_3^p: p_1 = p_3 = p_4$$
$$H_3^\mu: \mu_1 = \mu_3 = \mu_4$$

Donde p son las proporciones de respuesta y μ son las medias de cada cluster.

Gráfico 26 - Segmentación experimento



Fuente: Elaboración propia

7.5.4. Tamaños muestrales

Para el cálculo de los tamaños muestrales es importante definir si el test de hipótesis es de una cola o de dos, la diferencia se basa en si se sabe a priori cuál de las dos tasas que se está comparando es mayor que la otra. Es importante notar que para las hipótesis 1 y 3 sí se puede afirmar que unas tasas deben se mayor que otras, por lo que el test debe ser de una cola. Para el experimento 2 no se puede afirmar, por lo que el test, y el tamaño muestral, debe ser calculado para dos colas.

Para todos los efectos se considera un nivel del confianza del 90% y un poder de la muestra del 80%. Así, $Z_\alpha = 1{,}282$ para una cola y $Z_\alpha = 1{,}645$ para dos colas, mientras que $Z_\beta = 0{,}842$.

Para la estimación de las tasas de respuesta se analizaron las tasas de respuesta de las campañas de las que se tenía información. Además, se aumento en 10% la cantidad necesaria dada y se redoendeó. Así, los tamaños muestrales para los distintos experimentos son:

- Para evaluar la hipótesis 1 se debe realizar un test de proporciones de una cola, para un esperando que el grupo detectado por el RandomForest tenga una tasa de respuesta del 6% y el aleatorio del 3%. Se requiere que cada grupo sea de 430 personas, para dejar un margen se recomiendan 500 personas por grupo. Tanto el grupo aleatorio como el grupo RandomForest debe tener un grupo de control. Es importnte

realizar el mismo estímulo para ambos grupos y no realizar estímulo alguno a los grupos de control.

- Para evaluar la hipótesis 2 se debe realizar un test de proporciones y uno de medias, en este caso de dos colas. Para el test de proporciones, considerando tasas de respuesta de 6% y 3%, se requieren 590 personas. Para el test de medias, considerando una desviación estándar de 0,24 (desviación estándar observada en el total de los datos), queriendo considerar diferencias de medias de al menos 3 puntos porcentuales, la cantidad requerida por grupo es 8254. Por lo anterior se recomienda 9100 personas más grupo de control, por cada segmento a testear.

- Para evaluar la hipótesis 3 se debe realizar un test de proporciones y un test de medias de una cola. Para el test de proporciones se pueden usar el valor calculado para la hipótesis 1, es decir, 500 individuos por cluster. Para el test de medias, considerando una desviación estándar de 0,24 y diferencias de medias de al menos 3 puntos porcentuales, la cantidad de personas requeridas es 2478 por grupo. Por lo anterior se recomiendan 2800 personas de cada cluster, más el grupo de control de cada cluster.

Es importante que, luego de la creación de las bases, se revise que las proporciones de distintas variables sea similar en los distintos grupos, y no haya quedado algún grupo desbalanceado, por ejemplo, en la edad con respecto al grupo de control. Además, se recomienda que los grupos de control no sean ni muy grandes ni muy pequeños, es usual usar un tamaño del 10% del grupo original.

7.5.5. Evaluación

Lo primero que se debe realizar es el diseño del estimulo, se puede trabajar con una campaña de aumentos genérica o enmarcada en alguna festividad, además, el estímulo puede ser directo o ir inserto dentro de otra comunicación. A modo de ejemplo se presenta en la Ilustración 15 una campaña genérica directa, en la que el objetivo de la comunicación es comunicarle la campaña de aumentos al socio. En la Ilustración 16 se muestra una campaña genérica inserta en otra comunicación.

La importancia de esta diferencia es que las campañas directas tienen el asunto del correo relacionado con el contenido de la campaña. Por el contrario, cuando la campaña que va inserta en otra comunicación, el asunto del correo está relacionado con la comunicación principal. Así, para los ejemplos presentados anteriormente, un ejemplo de asunto de mensaje podría ser "Necesitamos 5.000 aumentos de aporte! Súmate!" para la campaña directa y "Feliz 2020 te desean los abuelos! Revisa nuestro 2019" para la campaña inmersa.

Ilustración 15 - Campaña genérica directa

Fuente: Elaboración propia

Ilustración 16 - Campaña genérica inmersa

Fuente: Elaboración propia

La tasa de lectura o apertura (TL) y tasa de respuesta (TR) son dos tasas que se utilizan comunmente para la evaluación de las campañas. La TL, al ser una medida de la proporción de socios que abrieron la comunicación, está considerando como estímulo sólo el asunto del mensaje, ya que es lo único que ha visto el socio hasta antes de abrir el mensaje. En el caso en que la campaña va inserta en otra comunicación, las personas que no abrieron el correo nunca recibieron la campaña, por lo que en ese caso sólo se puede considerar la TR para evaluar la campaña.

Las plataformas para manejo de correos masivos entregan la cantidad de correos recibidos, esto es, que no rebotaron, la cantidad que fueron abiertos y, se reciben y pueden registrar la cantidad de respuestas. Con lo anterior y luego de un tiempo en el que se reciben las respuestas, este periodo en general

está entre 3 semanas y un mes, se calculan las TL y TR y se procede a calcular los estadísticos comentados en el apartado 4.3.4, así saber si estas tasas son significativamente diferentes, por ejemplo entre un grupo de control y de tratamiento.

7.6. Confección Modelo Up-Lift

Como se explicó en el apartado 4.1.3, existen dos maneras de modelar un up-lift, la primera es modelar cada grupo por separado y luego restarlos. Otra forma es modelar directamente el up-lift de la forma:

$$lift = P(y = 1|tratamiento = 1) - P(y = 1|tratamiento = 0)$$

De esta forma, se obtiene un medida más sincera del up-lift.

Para modelar el up-lift se debe crear una nueva variable binaria con los individuos a los que se les aplicó el tratamiento. Al no tener los resultados de los experimentos, a modo de ejemplo, se reemplazó por una variable arbitraria, por lo que los resultados a continuación no son reales, solo se incluyen a modo de ejemplo para mostrar cómo es el análisis que debe hacerse una vez se hayan realizado los experimentos.

El modelo Up-Lift trabaja con distintas medidas de divergencia, esto es, para separar a los grupos. Estas medidas se pueden comparar con el criterio de Qini. Este criterio obtiene la ganancia de información, por lo que se debe escoger el que mayor valor tenga. A continuación, en la Tabla 11, se presentan los métodos de divergencia con sus valores de Qini.

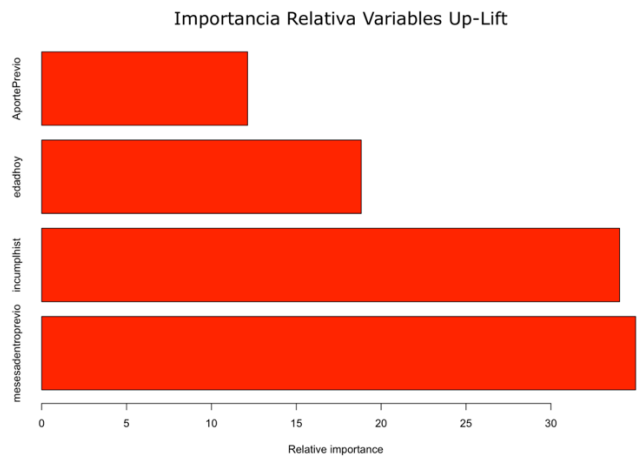
Tabla 11 - Q de Quini

	Divergencia Kullback-Leibler	Divergencia Euclidiana	Divergencia Chi-Cuadrado	Método de Iteraciones
Q de Qini	0,0305	0,0305	0,000923	0,0662

Fuente: Elaboración propia

Se puede ver que el criterio de divergencia con mayor índice de Qini es el método de iteraciones.

Gráfico 27 - Importancia Relativa Up-Lift

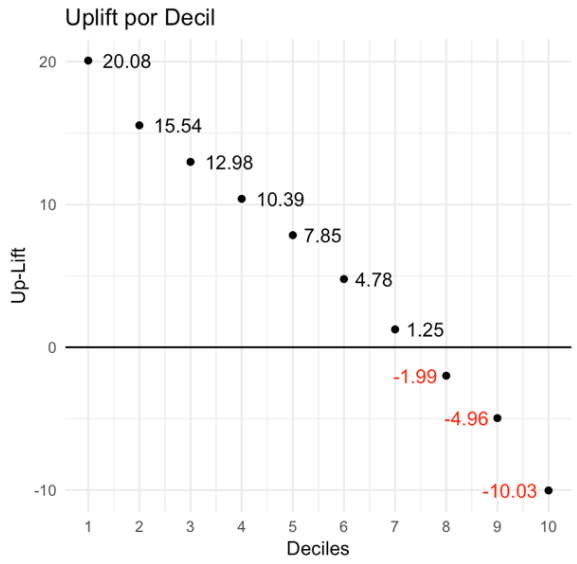


Fuente: Elaboración propia

Una vez escogido el criterio de divergencia, se procede a entrenar el modelo, al igual que para RF y Logit, se separa una porción de la base para entrenamiento y otra para prueba. Como resultado entrega la importancia relativa de las variables (ver Gráfico 27). Al igual que en la construcción de otros modelos, se puede ir agregando o eliminando variables según su importancia relativa hasta llegar al mejor modelo.

Además se obtiene el lift de cada persona, así, se pueden ordenar y crear deciles. Gráficamente se debise obtener algo similar al Gráfico 28.

Gráfico 28 - Up-Lift por decil



Fuente: Elaboración propia

Se puede ver en este gráfico cómo se comportan los deciles, analizándo los incrementos en el lift, los deciles 1 al 4 serían los influenciables, por tener un up-lift superior al 10%. Luego, muy poco por sobre el 0, los deciles 5 al 7 serían los seguros, ellos no necesitan ser influenciados, van a comprar. Luego los deciles 8 y 9 serían los resistentes, no comprarán aunque se les trate. Finalmente los no-molestar, serían los pertenecientes al decil 10.

8. Conclusiones y Recomendaciones

Inicialmente se tuvieron dificultades con la base de datos, esta no tiene muchas variables de comportamiento y estructuralmente está muy sucia. Pese a ello, se logró estructurar y crear modelos y hacer análisis al respecto.

Esta condición de la base se debe a que no existe una estrategia que apunte hacia la generación de datos, esto es esencial para poder realizar estrategias de marketing directo. Se recomienda a la institución apuntar en esa dirección en nivel de datos, ya que poseen información que no están digitalizando o no están utilizando. Se pueden obtener variables que expliquen el comportamiento y la participación de las personas de otras maneras con FLR, por ejemplo en la participación de la colecta o si está en alguna de las redes sociales.

Se logró desarrollar un modelo predictivo, considerando que existen las capacidades para la realización de experimentos en la institución, se podría testear. Dada la poca data, los experimentos son una muy buena fuente de información para campañas con marketing directo, así, se puede levantar heterogeneidad.

Los modelos que se desarrollaron en esta memoria no son de gran complejidad ni de gran costo, fueron procesado a traves del programa R que es de uso gratuito, por lo que replicar estudios similares no debiese representar un gran costo para la institución, así poder medir distintas variables de los suscritos al programa y rentabilizar cada vez más la cartera de socios.

Finalmente, se encontró que existe una relación directa entre la antigüedad de las personas en el programa y su predisposición a realizar un aumento. Además, se encontró que existe una relación cuadrática con el aporte, llevado a la realidad, esto quiere decir que a mayor aporte mayor predisposición a aceptar un aumento, hasta cierto límite, en que si se traspasa, esta predisposición comienza a descender.

De los modelos de regresión numérica y logística se pudo apreciar la diferencia en costo de estimar un valor versus en estimar si se realizará o no la acción. Es menos costoso predecir quiénes se activarán o no a predecir en cuánto se activarán. A nivel de experimentos se aprecia la diferencia con las distintas cantidades que se requieren para un test de medias o de proporciones, siendo más costoso el de medias.

Otra manera de controlar las campañas comunicacionales es a través de la tasa de fuga, tema que está fuera del alcance de la memoria, esta tasa puede ir variando conforme a las campañas, especialmente en los deciles de los no-molestar. Se pueen realizar trabajos similares a este pero considerando la tasa de fuga en lugar de la tasa de activación.

Bibliografía

[1] Fundación las Rosas. Fundación las Rosas. [en línea] <http://www.fundacionlasrosas.cl> [consulta: 13 nov.-18]

[2] Fundación las Rosas. Ingreso y hogares. [en línea] <http://www.fundacionlasrosas.cl/ingreso-y-hogar/> [consulta: 13 nov.-18]

[3] Fundación las Rosas. Nosotros. [en línea] <http://www.fundacionlasrosas.cl/nosotros/> [consulta: 10 abril.-19]

[4] Fundación las Rosas. Informe de Gestión Web, al mes de abril 2018. [en línea] <http://www.fundacionlasrosas.cl/informes_transparencia/IDG/Propuesta_2018.pdf> [consulta: 13 nov.-18]

[5] Fundación las Rosas. Informe de Gestión Web, al mes de diciembre 2018. [en línea] <http://www.fundacionlasrosas.cl/informes_transparencia/IDG/Propuesta_Informe_Gestion_Dic_2018.pdf > [consulta: 11 jun.-19]

[6] Gerencia Fundación las Rosas de Ayuda Fraterna. Presentación *Cómo vamos Noviembre 2018*. [archivo interno]

[7] SENAMA. Catastro ELEAM. [en línea] <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwjeto6QyNzcAhUMGJAKHRlAD3UQFjABegQICRAC&url=http%3A%2F%2Fcatastroeleam.senama.cl%2Fexportar.php&usg=AOvVaw13agEwT41l_PxApuJxaDJl> [consulta: 22 nov.-18]

[8] Biblioteca del Congreso Nacional de Chile. DTO14 05-AGO-2010 Ministerio de Salud, Subsecretaría de Salud Pública. [en línea] <https://www.leychile.cl/Navegar?idNorma=1015936&idVersion=2011-10-04> [consulta: 22 nov.-18]

[9] Servicio de Impuestos Internos. Donaciones. [en línea] <http://www.sii.cl/contribuyentes/actividades_especiales/donaciones.htm> [consulta: 22 nov.-18]

[10] Servicio de Impuestos Internos. Ley sobre impuestos a la Herencia, Asignaciones y Donaciones. [en línea] <http://www.sii.cl/pagina/jurisprudencia/legislacion/basica/herencias_asignaciones_donaciones.htm> [consulta: 22 nov.-18]

[11] Dirección del Trabajo. Código del Trabajo [en línea]

<http://www.dt.gob.cl/portal/1626/articles-95516_recurso_2.pdf> [consulta: 22 nov.-18]

[12] Dirección de Administración y Proyectos, Fundación las Rosas de Ayuda Fraterna. Mayor contable. [archivo interno]

[13] Cadem. III Versión Estudio de Marcas Ciudadanas. [en línea] <https://www.cadem.cl/wp-content/uploads/2018/08/Marcas-ciudadanas-2018-VF-Presentación.pdf> [consulta: 22 nov.-18]

[14] Instituto Nacional de Estadísticas. Síntesis de Resultados Censo 2017. [en línea] <http://www.censo2017.cl/descargas/home/sintesis-de-resultados-censo2017.pdf> [consulta: 19 dic.-18]

[15] Chile Atiende. Pensión Básica Solidaria de Vejez (PBSV). [en línea] < https://www.chileatiende.gob.cl/fichas/5270-pension-basica-solidaria-de-vejez-pbsv > [consulta: 19 dic.-18]

[16] Ministerio de Desarrollo Social. Situación de Probreza. CASEN 2017. [en línea] <http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/docs/Resultados_pobreza_Casen_2017.pdf> [consulta: 19 dic.-18]

[17] Puro Marketing. El efecto Halo y su aplicación en el mundo del marketing. [en línea] <https://www.puromarketing.com/27/25850/efecto-halo-aplicacion-mundo-marketing.html> [consulta: 20 de dic.-18]

[18] Aguirre San Martín, 2017. PERFILAMIENTO DE CLIENTES INFLUENCIABLES EN CAMPAÑAS DE PRODUCTOS FINANCIEROS EN UNA EMPRESA DE RETAIL FINANCIERO. Universidad de Chile. Santiago

[19] Porras Ignacio, 2011. SEGMENTACIÓN DE CLIENTES DE UNA EMPRESA MULTISERVICIO PARA LA RENTABILIZACIÓN DE SU CARTERA. Universidad de Chile. Santiago

[20] Anibal Goicochea. CRISP-DM, Una metodología para proyectos de Minería de Datos. [en línea] < https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/> [consulta: 24 dic.-18]

[21] Rodriguez Oldemar. Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. Universidad Nacional de Costa Rica.

[22] Reus Canal Díaz, Comparción de proporciones, Revista Seden. [en línea] < http://www.revistaseden.org/files/11-CAP%2011.pdf > [consulta: 11 abril -19]

[23] Jorge Dagnino S., Análisis de Proporciones, Revista Chilena de Anestesia. [en línea]
<http://revistachilenadeanestesia.cl/PII/revchilanestv43n02.12.pdf>
[consulta: 11 abril -19]

[24] Argüello, Pinzón, (2016). El modelo de up-lifting aplicado a un score de riesgo. Universidad Santo Tomás, Bogotá, Colombia.

[25] Rzepakwski and Jaroszewicz, (2012). Uplift Modeling in Direct Marketing, National Institute of Telecommunications, Warsaw, Poland.

[26] R. Mora, El Modelo Logit y el Modelo Probit. Departamento de Economía, Universidad Carlos III de Madrid. [en línea]
<http://www.eco.uc3m.es/~ricmora/miccua/materials/S07T21_Spanish_handout.pdf> [consulta: 11 abril -19]

[27] Minería de Datos. Grupo de Investigación MIDAS [en línea]
<
https://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf > [consulta: 3 sept. -19]

[28] El proceso KDD, Big Data pata todos. [en línea]
< http://traduccionesbigdata.blogspot.com/2017/07/el-proceso-kdd.html>
[consulta: 11 abril -19]

[29] Ryan Zhao, (2012). Improve Marketing Campaign ROI using Uplift Modeling. Analytics Resourcing Centre.

[30] Descripción funcional – CareCloudExtern. [en línea]
< https://sites.google.com/a/grupoica.com/carecloudextern/descripcion-funcional> [consulta: 30 ago -19]

[31] Álvarez, Rafael, Estadística multivariante y no paramétrica con SPSS: aplicación a las ciencias de la salud, Madrid: Ediciones Díaz de Santos, 1994.

[32] Colin Cameron, A., Windmeijer, Frank A.G., Gramajo, H., Cane, D., Khosla, C., "An R-squared measure of goodness of fit for some common nonlinear regression models", Journal of Econometrics, 1997.

[33] Olivares, M., Regresión Lineal, Estadística para la Economía y Gestión. [en línea]
< https://www.u-cursos.cl/ingenieria/2013/2/IN3401/2/material_docente/bajar?id_material=779135 > [consulta: 3 sept. -19]

[34] Hands-on Machine Learning with R, Chapter 11 Random Forest. [en línea] <https://bradleyboehmke.github.io/HOML/random-forest.html> [consulta: 3 sept. -19]

[35] MAYSA Consultores, Marketing Directo. [en línea]
<http://maysaconsultores.com.ar/recursos/Brochure_uplift.pdf> [consulta: 3 sept. -19]

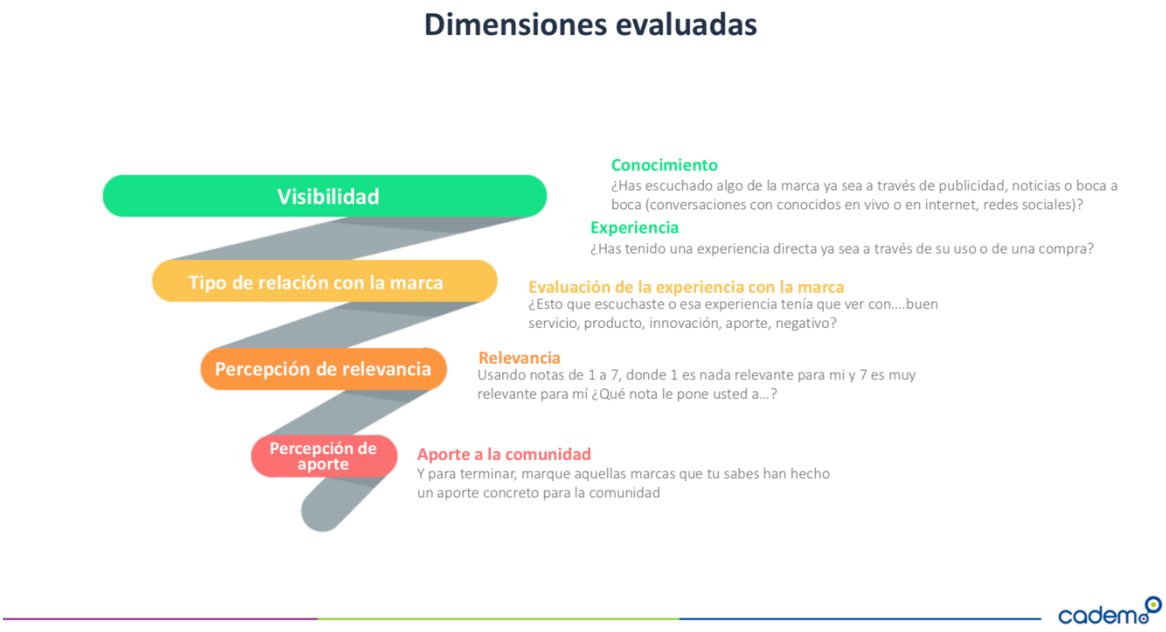
Anexos

Tabla 12 - Egresos por Ítem

Etiquetas de fila	Suma de REAL AÑO 2015	Suma de REAL AÑO 2016	Suma de REAL AÑO 2017	Suma de REAL AÑO 2018
1. REMUNERACIONES	**-527.734**	**-515.629**	**-574.040**	**-594.252**
ASESORIAS	-2.318	-2.823	-2.565	-1.484
BENEFICIOS	-15.329	-15.844	-23.579	-19.152
BONO ASISTENCIA	-6.057	-5.229	-4.722	-3.702
CAPACITACION	-499	-738	-1.269	-1.024
COMISIONES	-16.512	-17.350	-17.479	-15.758
FINIQUITOS	-29.362	-19.264	-24.932	-17.125
HONORARIOS	-3.191	-3.408	-3.849	-3.850
HORAS EXTRAS	-5.166	-4.482	-4.958	-4.850
LEYES SOCIALES	-12.378	-12.145	-14.874	-14.593
OTRAS REMUNERACIONES	-4.674	-3.500	-4.019	-3.787
PREVENCION DE RIESGOS	-1.721	-1.909	-1.914	-2.504
REMUNERACIONES SENAMA	-4.234	-4.800	-6.081	-6.192
SUELDOS Y SALARIOS	-426.292	-424.136	-463.800	-500.231
10. PROMOCION	**-1.842**	**-2.023**	**-2.648**	**-2.624**
PUBLICIDAD	-1.842	-2.023	-2.648	-2.624
11. SERV GENERALES Y ASEO	**-9.742**	**-10.598**	**-13.099**	**-12.219**
DETERGENTES	-1.078	-1.086	-1.279	-1.538
OTROS SERV GENERALES	-2.233	-2.062	-2.553	-1.891
SERVICIOS EXTERNOS	-3.461	-5.098	-6.013	-5.713
UTILES DE ASEO	-2.970	-2.352	-3.254	-3.077
12. ARTICULOS DE OFICINA	**-1.344**	**-1.344**	**-1.751**	**-1.541**
INSUMOS TECNOLOGICOS	-300	-160	-329	-294
OTROS ART. OFICINA	-1.015	-1.175	-1.393	-1.219
PAPELERIA	-29	-9	-30	-29
13. ARRIENDO DE INMUEBLE	**-3.308**	**-3.738**	**-3.851**	**-3.772**
ARIENDO OFICINAS Y LOCALES	-3.308	-3.738	-3.851	-3.772
14. LEGAL E IMPUESTOS	**-1.204**	**-1.628**	**-1.591**	**-1.723**
BIENES RAICES	-516	-957	-994	-1.027
VARIOS	-688	-671	-596	-696
15. TRANSPORTES Y VIAJES	**-5.462**	**-6.326**	**-7.735**	**-6.592**
FLETES	-525	-702	-895	-230
MOVILIZACION EXTERNA	-1.442	-1.474	-2.704	-2.577
VEHICULOS	-2.189	-2.615	-2.164	-2.062
VIAJES Y ESTADIA	-1.306	-1.535	-1.972	-1.723
16. SEGUROS	**-996**	**-1.052**	**-1.194**	**-1.305**
BIENES MUEBLES	-522	-468	-598	-584
OTROS SEGUROS	-237	-376	-408	-473
VEHICULOS	-237	-208	-187	-248
17. AYUDA Y DONACION	**-3.709**	**-3.286**	**-3.264**	**-3.246**
AYUDA Y DONACIONES	-1.941	-1.395	-1.223	-1.173
DEVOLUCION AL RESIDENTE	-1.768	-1.891	-2.041	-2.073
18. TEMPLOS Y CAPILLAS	**-874**	**-1.163**	**-1.138**	**-464**
TEMPLO Y CAPILLAS	-874	-1.163	-1.138	-464
19. VARIOS	**-1.799**	**-4.425**	**-2.549**	**-1.586**
SUSCRIPCIONES Y ACT EXTRAPROGRAMATICAS	-1.799	-4.425	-2.549	-1.586
2. COMISIONES BANCARIAS	**-13.288**	**-11.468**	**-10.000**	**-10.416**
COMISION ADMINISTRATIVA	-82	-60	-5	0
COMISION PAC	-10.513	-10.695	-9.213	-9.395
COMISION PAT	-380	-293	-503	-513
OTRAS COMISIONES	-2.314	-420	-279	-509
3. SERVICIOS BASICOS	**-43.537**	**-43.092**	**-49.519**	**-43.432**
AGUA, LUZ Y GAS	-35.325	-34.497	-39.914	-35.683
LEÑA Y COMBUSTIBLES	-1.370	-1.212	-2.286	-2.065
TELEFONIA	-6.842	-7.383	-7.319	-5.684
4. ALIMENTACION	**-37.271**	**-34.648**	**-34.692**	**-33.502**
ALIMENTOS	-32.764	-29.244	-29.149	-27.341
ALMUERZOS	-4.507	-5.405	-5.543	-6.161
5. MANTENCION	**-11.014**	**-13.384**	**-23.899**	**-14.839**
FUMIGACIONES	-974	-982	-909	-898
MANT. DE INMUEBLE	-1.850	-2.507	-6.425	-4.004
MANT. MAQUINAS Y EQUIPOS	-3.648	-3.309	-3.074	-2.860
MANT. SISTEMAS BÁSICOS	-2.091	-2.225	-3.883	-3.100
MANT. TECNOLÓGICA	-310	-1.200	-540	-449
MANT. VEHICULOS	-798	-973	-1.177	-1.010
REPARACIONES	-1.343	-2.188	-7.891	-2.518
6. SALUD	**-13.090**	**-15.236**	**-12.888**	**-10.772**
INSUMOS MEDICOS	-2.386	-3.288	-2.645	-2.141
MEDICAMENTOS	-4.191	-5.439	-3.851	-3.555
OTROS GASTOS DE SALUD	-4.101	-4.341	-3.142	-2.019
SERVICIOS CLÍNICOS	-2.412	-2.168	-3.250	-3.058
7. INSUMOS DE CUIDADO	**-15.032**	**-12.035**	**-19.646**	**-20.138**
GUANTES	0	0	-405	-1.631
OTROS INSUMOS	-2.942	-841	-2.818	-523
PAÑALES	-12.090	-11.194	-16.423	-17.984
8. INSUMOS HOGAR Y OTROS	**-1.523**	**-1.966**	**-3.295**	**-3.818**
BIENES NO INVENTARIABLES	0	0	-1.049	-386
GASTOS VARIOS HOGAR	-402	-205	-498	-283
GTOS. RECURSOS HOGARES	-1.121	-1.761	-1.748	-3.149
9. COSTO DE VENTAS	**-3.251**	**-6.700**	**-10.404**	**-10.402**
COSTO BAZARES	-941	-943	-958	-1.152
COSTO DE CAMPAÑAS	-450	-3.027	-5.456	-6.581
COSTO DE EVENTO Y PAVO	-1.704	-2.326	-2.588	-2.462
COSTO DE PRODUCTOS	-156	-405	-1.402	-207
Total general	**-696.021**	**-689.739**	**-777.202**	**-776.642**

Fuente: Elaboración propia [12]

Ilustración 17 - Dimensiones Evaluadas Cadem



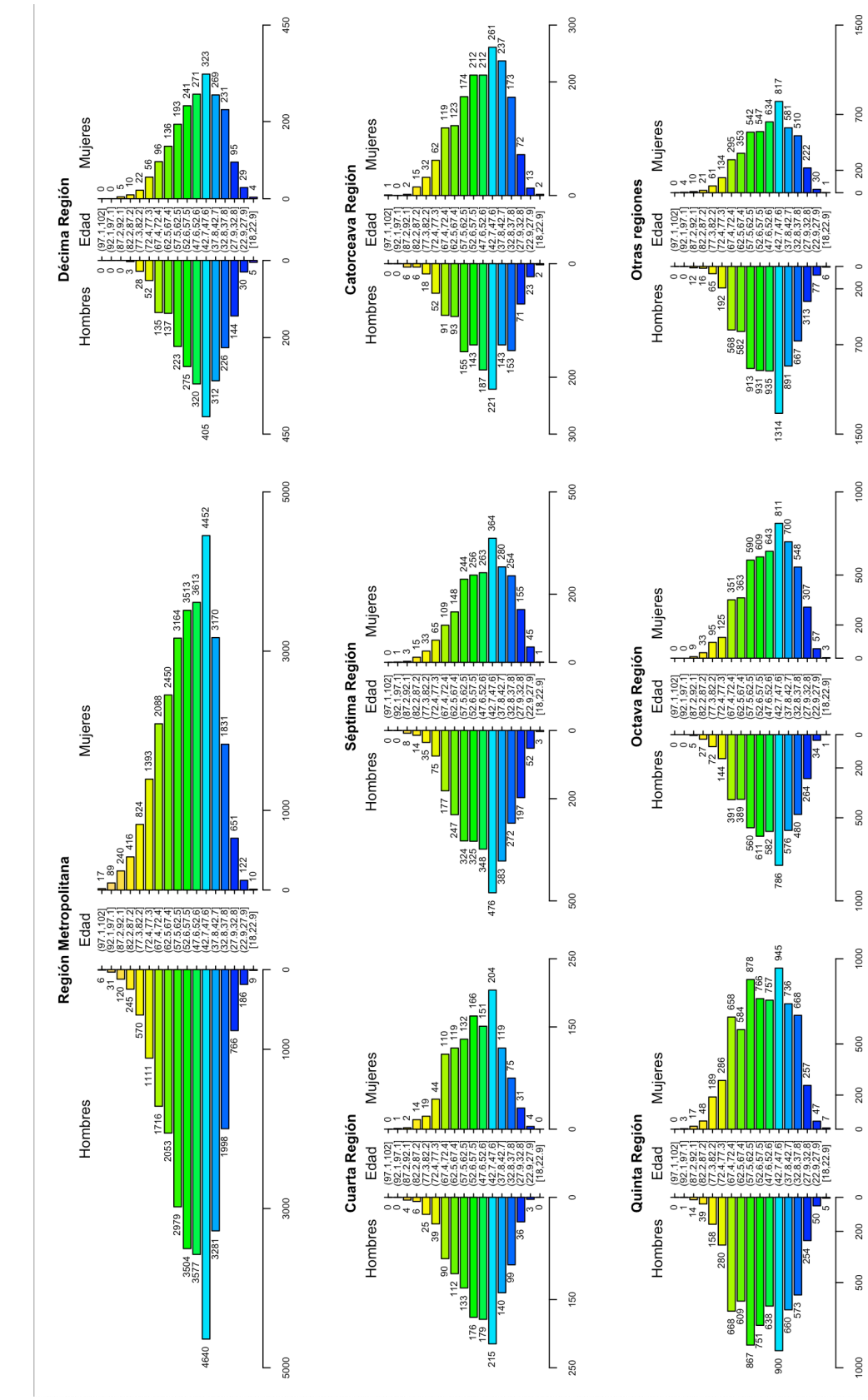
Fuente: [13]

Ilustración 18 - Percentiles variable AportePesos

```
> quantile(basegigante$AportePesos, prob = seq(0, 1, length = 101))
     0%      1%      2%      3%      4%      5%      6%      7%      8%      9%     10%     11%
  100.0  2000.0  2500.0  3000.0  3000.0  3000.0  3000.0  3000.0  3000.0  3000.0  3000.0  3500.0
    12%     13%     14%     15%     16%     17%     18%     19%     20%     21%     22%     23%
 3500.0  3500.0  3500.0  3500.0  3500.0  3500.0  3500.0  3500.0  3500.0  4000.0  4000.0  4000.0
    24%     25%     26%     27%     28%     29%     30%     31%     32%     33%     34%     35%
 4000.0  4000.0  4000.0  4000.0  4000.0  4500.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0
    36%     37%     38%     39%     40%     41%     42%     43%     44%     45%     46%     47%
 5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0
    48%     49%     50%     51%     52%     53%     54%     55%     56%     57%     58%     59%
 5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5000.0  5530.0  6000.0  6000.0  6000.0
    60%     61%     62%     63%     64%     65%     66%     67%     68%     69%     70%     71%
 6000.0  6000.0  7000.0  7000.0  7000.0  7000.0  7000.0  7000.0  7000.0  7500.0  7500.0  7500.0
    72%     73%     74%     75%     76%     77%     78%     79%     80%     81%     82%     83%
 8000.0  8000.0  8000.0  8000.0 10000.0 10000.0 10000.0 10000.0 10000.0 10000.0 10000.0 10000.0
    84%     85%     86%     87%     88%     89%     90%     91%     92%     93%     94%     95%
10000.0 10000.0 10000.0 10000.0 10500.0 10500.0 10500.0 12000.0 12500.0 13000.0 15000.0 15000.0
    96%     97%     98%     99%    100%
15000.0 20000.0 20000.0 26622.8 500000.0
```

Fuente: Elaboración propia

Gráfico 29 - Cantidad de Amigos según sexo y edad, por región



Fuente: Elaboración propia

Tabla 13 - Escalado estandar vs Escalado MINMAX

<table>
<tr><th></th><th>Estalado estándar</th><th>Escalado MINMAX</th></tr>
<tr><td>Regresión Lineal</td><td>

```
Call:
lm(formula = porc_ult_aum ~ Genero + AportePrevioN + mesesadentroN +
    edadhoyN + incumplhist, data = basegigante30k)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1362 -0.0840 -0.0589 -0.0246  9.4506

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0706582  0.0010615  66.565  < 2e-16 ***
GeneroM        0.0084536  0.0013405   6.306 2.87e-10 ***
AportePrevioN -0.0079310  0.0006853 -11.573  < 2e-16 ***
mesesadentroN  0.0125524  0.0007703  16.294  < 2e-16 ***
edadhoyN       0.0084486  0.0007140  11.832  < 2e-16 ***
incumplhist   -0.0910775  0.0033715 -27.014  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2397 on 128895 degrees of freedom
Multiple R-squared:  0.01879,   Adjusted R-squared:  0.01875
F-statistic: 493.7 on 5 and 128895 DF,  p-value: < 2.2e-16
```

</td><td>

```
Call:
lm(formula = porc_ult_aum ~ Genero + AportePrevio + mesesadentro +
    edadhoy + incumplhist, data = basegigante30k)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1362 -0.0840 -0.0589 -0.0246  9.4506

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.043765   0.002511  17.427  < 2e-16 ***
GeneroM       0.008454   0.001341   6.306 2.87e-10 ***
AportePrevio -0.060884   0.005261 -11.573  < 2e-16 ***
mesesadentro  0.063318   0.003886  16.294  < 2e-16 ***
edadhoy       0.053482   0.004520  11.832  < 2e-16 ***
incumplhist  -0.091077   0.003371 -27.014  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2397 on 128895 degrees of freedom
Multiple R-squared:  0.01879,   Adjusted R-squared:  0.01875
F-statistic: 493.7 on 5 and 128895 DF,  p-value: < 2.2e-16
```

</td></tr>
<tr><td>Regresión Logística</td><td>

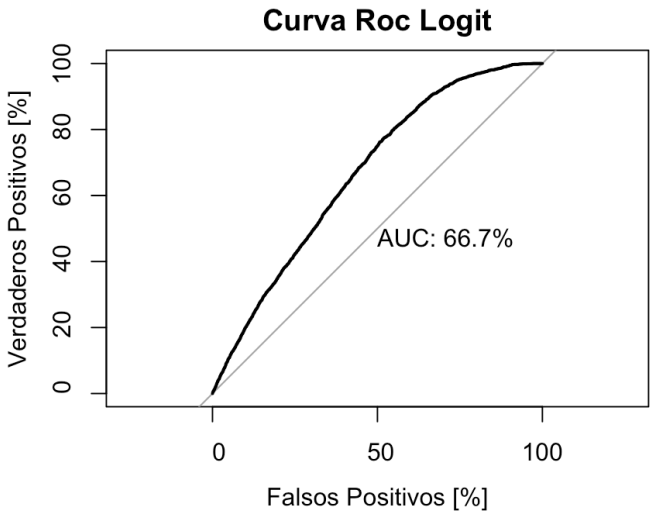
```
Call:
glm(formula = binaria ~ Genero + AportePrevioN + mesesadentroN +
    edadhoyN + incumplhist, family = binomial, data = basegigante30k)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8298  -0.5724  -0.5073  -0.2579   3.5213

Coefficients:
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)   -1.676609   0.014567 -115.098  < 2e-16 ***
GeneroM        0.073154   0.017657    4.143 3.43e-05 ***
AportePrevioN  0.038051   0.008745    4.351 1.35e-05 ***
mesesadentroN -0.028693   0.010449   -2.746  0.00603 **
edadhoyN       0.177509   0.009236   19.218  < 2e-16 ***
incumplhist   -4.723983   0.110245  -42.850  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93198  on 128900  degrees of freedom
Residual deviance: 88129  on 128895  degrees of freedom
AIC: 88141

Number of Fisher Scoring iterations: 6
```

</td><td>

```
Call:
glm(formula = binaria ~ Genero + AportePrevio + mesesadentro +
    edadhoy + incumplhist, family = binomial, data = basegigante30k)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8298  -0.5724  -0.5073  -0.2579   3.5213

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.13594    0.03493 -61.149  < 2e-16 ***
GeneroM       0.07315    0.01766   4.143 3.43e-05 ***
AportePrevio  0.29211    0.06713   4.351 1.35e-05 ***
mesesadentro -0.14474    0.05271  -2.746  0.00603 **
edadhoy       1.12367    0.05847  19.218  < 2e-16 ***
incumplhist  -4.72398    0.11024 -42.850  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93198  on 128900  degrees of freedom
Residual deviance: 88129  on 128895  degrees of freedom
AIC: 88141

Number of Fisher Scoring iterations: 6
```

</td></tr>
</table>

Gráfico 30 - Curva ROC Logit AportePesos^2



Fuente: Elaboración propia