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

IDENTIFICACIÓN DE CLIENTES DE MAYOR RELEVANCIA MEDIANTE TEORÍA DE GRAFOS CON FOCO EN CAMPAÑAS DE APERTURA DE TARJETA DE CREDITO PARA UN BANCO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JOSÉ FELIPE SÁNCHEZ VEGA

PROFESOR GUÍA:
JUAN PABLO ROMERO GODOY

MIEMBROS DE LA COMISIÓN:
PABLO MARIN VICUÑA
NICOLÁS FRITIS COFRÉ

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: JOSÉ FELIPE SÁNCHEZ VEGA
FECHA: JUNIO 2020
PROFESOR GUIA: JUAN PABLO ROMERO

IDENTIFICACIÓN DE CLIENTES DE MAYOR RELEVANCIA MEDIANTE TEORÍA DE GRAFOS CON FOCO EN CAMPAÑAS DE APERTURA DE TARJETA DE CREDITO PARA UN BANCO

La eficiencia de las campañas para captar nuevos clientes en la industria financiera es de gran importancia para las empresas. Un mercado con gran competencia y con poca variabilidad de productos, obligan a buscar estrategias innovadoras que permitan mejorar la rentabilidad de sus productos y servicios.

En este proyecto, se busca generar una herramienta que permita identificar los clientes más relevantes para el banco en cuanto a su potencial de apertura de tarjeta de crédito utilizando algoritmos de grafos complementados con modelos de machine learning.

Las hipótesis se basan en la factibilidad de construir redes utilizando datos de transferencias de dinero, emitidas o recibidas por clientes del banco. A la vez, que es posible contrastar la relevancia de ellos utilizando campañas de apertura, en particular, campañas de referidos.

Un grafo se compone por nodos y enlaces, los que en este caso representan a personas (clientes) y sus relaciones (transferencias bancarias). Luego, se establecen métricas que puedan determinar el grado de influencia de una persona sobre una red y métricas que indiquen la calidad de los enlaces o relaciones.

El desarrollo se divide en la constitución de dos redes de clientes. Una primera red se construye utilizando datos de transferencias de los años 2017 y 2018 contrastando las métricas obtenidas con una campaña de referidos realizada en abril de 2018. Una segunda red se construye con datos de transferencias de los años 2018 y 2019. Las métricas obtenidas son evaluadas en una campaña de referidos realizada como piloto de este proyecto durante diciembre de 2019.

Los resultados muestran que no es posible generar un modelo con la bondad de ajuste necesaria para cuantificar relevancia a nivel de cliente. Sin embargo, se logra caracterizar la relevancia a nivel agregado, donde se observa que el conjunto identificado como “más importante” posee una tasa de respuesta de 4 veces la del grupo de control. Este clúster representa el 25% del universo de clientes analizado, sin embargo, acapara el 35% de la tasa de respuesta.

El proyecto muestra un VAN positivo con beneficios potenciales importantes debido a la directa implementación en las distintas unidades de negocio pertenecientes al mismo holding.

AGRADECIMIENTOS

Al cierre de este proyecto culmina también mi etapa universitaria, ha sido un proceso largo, lleno de altos y bajos que terminan por enmarcan la persona que soy hoy.

Agradezco enormemente a todas las personas que se involucraron de alguna u otra forma en esta etapa, a los amigos del colegio, de la universidad y de la vida. Ya sea que hayamos compartido un trabajo, una tarde o unas cervezas de alguna u otra forma se involucraron en mi vida y estuvieron ahí.

Hago un especial agradecimiento a mi familia, a mi hermana Sofia y a mis padres Julia y Jose por su apoyo y compañía incondicional. A mis padres por siempre preocuparse de que no me faltara nada, por estar ahí siempre que lo necesite, por las largas horas de estudio cuando era un niño, por su inagotable paciencia a la hora de educarme, y por sobre todo les agradezco por mostrarme que no existe un límite para mis metas ya que nunca le pusieron un límite a mi imaginación.

A la vez, no puedo dejar de mencionar a mis abuelos quienes desde que nací me han entregado un cariño infinito siendo un pilar muy importante en mi formación y en mi vida. En especial a mi abuelo Jose Sanchez, su historia, esfuerzo y perseverancia mostrados a lo largo de su vida son para mí una inagotable fuente de inspiración.

Finalmente, agradezco el tiempo y apoyo de mis profesores Guia y Coguia Juan Pablo Romero y Pablo Marin. A los miembros de la oficina donde realice este trabajo que de forma directa o indirecta se involucraron en este proyecto, ya sea respondiendo cosas de la pega o simplemente compartiendo un momento de distracción. Cierro los agradecimientos con una especial mención a Roberto Jara, quien se involucró mucho más allá de lo que implica su labor como tutor. Gracias por las horas acompañándome en el desarrollo del proyecto y por desde el primer momento creer en mí y en mi trabajo.

TABLA DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE FÓRMULAS

1 ANTECEDENTES GENERALES O INTRODUCCIÓN

1.1 Características de la organización/empresa

El trabajo de título se realiza en un banco filial de un holding que posee destacada participación en la región. Cuenta con presencia, en sus distinto líneas de negocio, en varios países de américa como Argentina, Perú, Colombia, México y Chile. Dentro de Chile, las principales unidades de negocio son servicios financieros, tiendas por departamento, supermercados y Marketplace.

La red de sucursales del banco cuenta con distintos formatos de oficina, ubicados en todas las regiones del país, con un total de 199 sucursales a lo largo de Chile.

El banco cuenta con una cartera de productos financieros como avance de tarjeta, créditos de consumo, crédito hipotecario, crédito automotriz y tarjetas de débito y crédito. Es en esta última donde el banco se posiciona como uno de los principales emisores a nivel nacional. A continuación, se presentan los emisores de tarjeta de crédito del 2018.

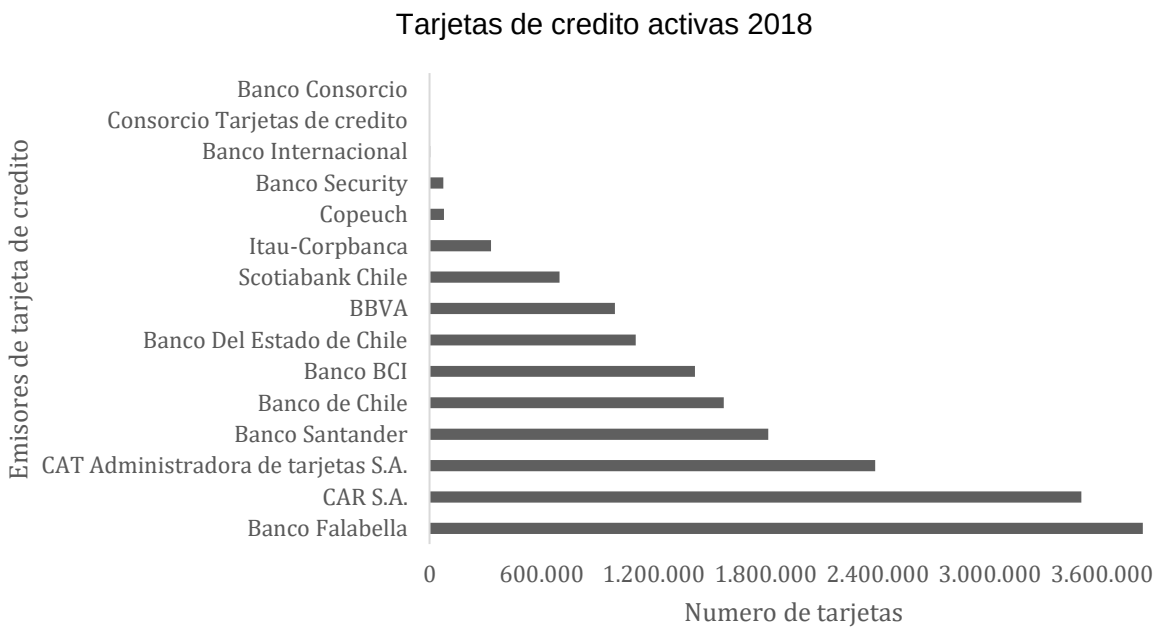


Ilustración 1: Tarjetas de crédito activas 2018

A partir de la ilustracion anterior, es posible evidenciar que en chile existen mas de 17 millones de tarjetas de credito activas. Esto equivale, en promedio, a 3 tarjetas por persona[1], haciendo de este un mercado altamente competitivo.

[1] https://www.biobiochile.cl/noticias/nacional/chile/2017/02/02/experto-en-finanzas-hay-mas-tarjetas-de-credito-en-chile-que-numeros-de-habitantes.shtml

1.2 Mercado y/o marco institucional

El sector industrial se define por instituciones financieras, esta categorización abarca un gran rango de instituciones que gestionan dinero, tales como, bancos, cooperativas, fondos de inversión, empresas del retail que ofrecen tarjetas de crédito y más. Por otro lado, existen varios organismos o cuerpos legales que regulan el quehacer de estas instituciones, siendo los principales:

- Superintendencia de Valores y Seguros (SVS)
- Comisión Para el Mercado Financiero (CMF) (Ex SBIF)
- Servicio Nacional del Consumidor (SERNAC)

1.3 Desempeño organizacional

Al estudiar la organización de la industria según colocaciones de consumo se distinguen 5 principales competidores con una participación superior al 12%. Esto muestra una industria concentrada donde estos exponentes abarcan un 75% del mercado. Las participaciones de mercado medidas por colocaciones de consumo total durante el 2018 se observan a continuación:

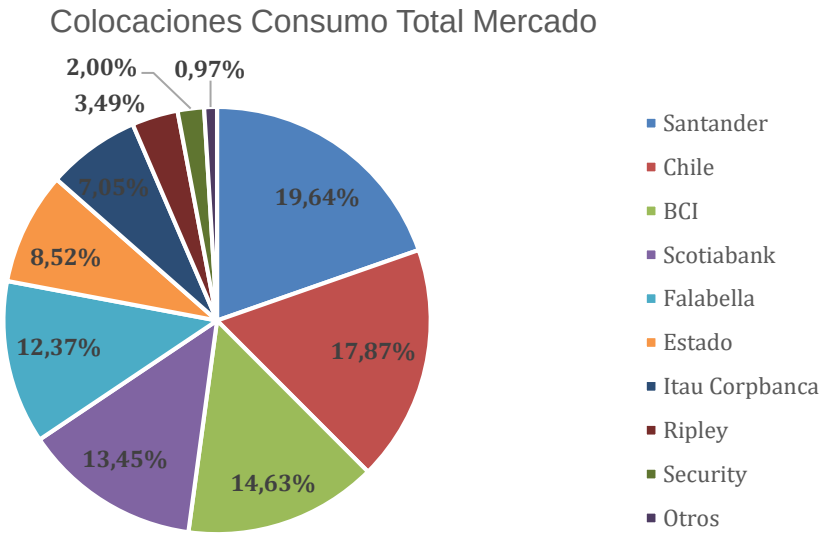


Ilustración 2: Participación de mercado bancario en Chile año 2018

Fuente: Elaboración propia datos CMF

El banco donde se desarrolla este proyecto tiene una participación de mercado superior al 10%. Esto se produce de forma reciente tras la fusión con una empresa del mismo holding que se hacía cargo de administrar su propia tarjeta de crédito. Este hito producido durante 2018 marca una fuerte tendencia en los análisis al desempeño histórico de la empresa.

Respecto al número de tarjetas de crédito del banco, se observa un crecimiento sostenido con un fuerte incremento en el 2018 por los motivos ya mencionados. Sin embargo, se observa un crecimiento sostenido desde el año 2013 al 2017 cercano al 10%.

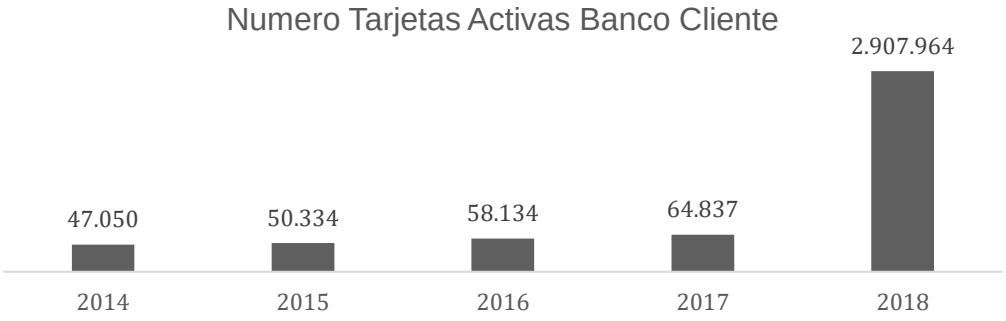


Ilustración 3: Tarjetas de crédito activas

Fuente: Información publica

2 DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

2.1 Información del área de la organización/empresa

El proyecto será abordado desde la gerencia de business intelligence la que tiene como objetivo encontrar insights a partir de las bases de datos del holding que permitan potenciar la toma de decisiones tanto a nivel directivo como operativo en las distintas unidades de negocio, y a la vez, potenciar las estrategias comerciales apalancadas en un entendimiento más profundo del cliente. Por lo tanto, se identifica como principales clientes internos de la gerencia a las áreas comerciales y de marketing.

El área se compone por 30 profesionales siendo la mayoría ingenieros civil industrial con foco en análisis de datos y machine learning, mientras que una pequeña parte corresponde a ingenieros civil en computación e ingeniería comercial.

El proyecto de memoria recae sobre el área de Machine Learning, bajo la subgerencia de Analytics y Machine Learning. El área está compuesta por 4 personas y tal como su nombre lo indica, tiene como objetivo acumular el conocimiento sobre esta materia para poder apoyar aplicando estas metodologías y algoritmos en los proyectos del resto de la gerencia.

Organigrama gerencia Business Inteligence

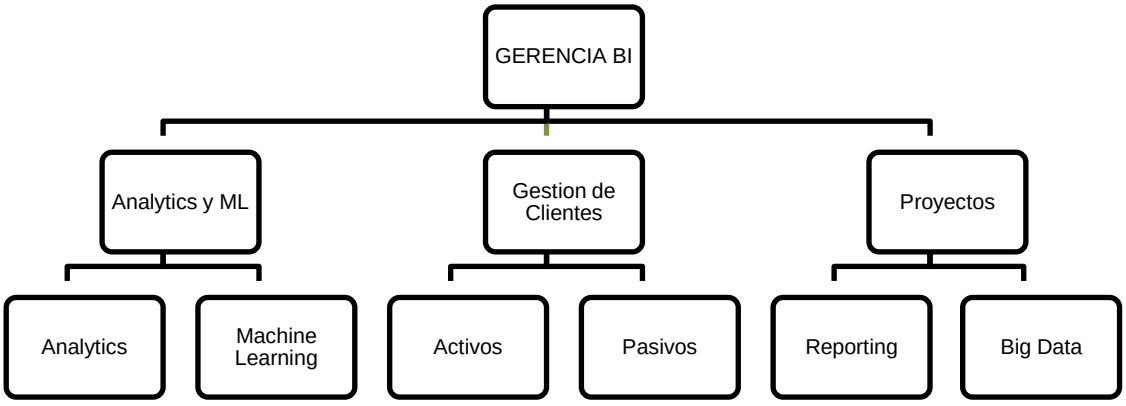


Ilustración 4: Estructura Organizacional Gerencia Business Intelligence

Para efectos de este proyecto se identifica como cliente al área de tarjetas. Esta tiene como objetivo gestionar las campañas de apertura de tarjeta de crédito del banco. Si bien el impacto del proyecto está más orientado al área de ánalytics, es en tarjetas donde los resultados del proyecto podrían ser aplicado de forma directa en el futuro.

2.2 Justificación

Este proyecto busca encontrar evidencia de la factibilidad y potenciales beneficios de la aplicación de modelamiento de redes sobre la base de clientes del banco.

Los análisis de redes tienen su fundamento sobre lo que en matemáticas se conoce como "Teoría de grafos". Los inicios de estos estudios se remontan al siglo XVIII de la mano de Leonar Euler quien plantea un primer problema de análisis de redes llamado "Bridges of Konisberg[2]".

En la actualidad, la masificación de internet, redes sociales y el desarrollo computacional en conjunto permiten aplicar modelamiento de redes sobre bases de relaciones cada vez más grandes, obteniendo métricas que permiten clasificar, perfilar e identificar a aquellas redes, nodos o clústers más relevantes.

2.2.1 Estado del arte

Para abordar los potenciales beneficios de un proyecto de redes en un banco resulta importante entender el estado del arte de esta materia.

Existen dos publicaciones que enmarcan el trabajo de este proyecto. En ambas se abordan las oportunidades de mejora y potenciales beneficios de aplicar un enfoque basado en redes a análisis de negocios que normalmente se realizan estudiando a cada cliente de forma particular.

[2] James D. Wilson, Data Instite conference 2017. University of San Francisco

Con el arribo de redes sociales y la masificación de los Marketplace online, es posible retener información valiosa sobre la influencia que pueden ejercer las personas sobre las decisiones de compra de sus pares. Como se demuestra en Mining the Network Value of Customers (Pedro Domingos, Matt Richardson, 2001). El efecto de esta "influencia" se puede materializar. En dicho estudio se analiza una base de clientes de un Marketplace de películas que ofrece a los clientes la opción de recomendar productos a sus contactos. Utilizando cadenas de markov logran aislar los ingresos esperados de la "influencia" que un cliente es capaz de ejercer sobre sus contactos.

En línea con lo anterior, surge la idea de valorizar a los clientes, no solo mediante la estimación de valor presente de los flujos de dinero derivados de sus compras futuras (enfoque clásico de Customer lifetime value), sino teniendo en consideración la influencia ejercida por un cliente sobre sus contactos y la que estos ejercen sobre el mismo.

La estimación del valor de vida de un cliente (CLV) puede ser de gran relevancia para una empresa. El valor de cada cliente, además de ser considerado un activo, se utiliza en la toma de importantes decisiones como la asignación de programas de fidelización, lo que se puede transformar en un factor crucial en el comportamiento de compra de un cliente en el futuro[3].

Como se muestra en Customer Lifetime Network Value (Julia Klier1, Mathias Klier1, Florian Probst, Lea Thiel1, 2014), existe evidencia que muestra que estudiar el valor de un cliente considerando su comportamiento de compra en el futuro de forma individual puede llevar a una sobre o sub estimación de su valor para la empresa. En dicho estudio se realiza una valoración de cada cliente bajo un enfoque de red denominada "Customer lifetime network value" (CLNV), la que incorpora efectos de influencia que emite y recibe un cliente. Los resultados son contrastados con el comportamiento de compra de los clientes de un OSN[4] (Online showtime network) mostrando que el CLNV mejora la estimación del valor de un cliente en comparación al clásico CLV.

2.2.2 Oportunidad

El estudio de redes representa una nueva forma de pensar y entender a los clientes y una gran oportunidad de encontrar valor sobre datos que no podrían ser explotados de otra manera y que no habían sido aprovechados.

Esta oportunidad se ve fuertemente potenciada por el contexto del banco donde se desarrolla el proyecto de memoria, el que lleva años dando énfasis en el análisis y procesamiento de datos para apalancar sus estrategias comerciales.

El gran número de clientes del banco en conjunto con la gran cantidad de información que se posee de ellos, permiten que sea el lugar idóneo para implementar análisis de redes.

[3] Fred Reichheld, Prescription for cutting cost. Bain & Company.
[4] Plataforma de streaming

Por un lado, se tiene que el número de clientes es un factor fundamental en esta materia ya que permite construir redes más concentradas que se traducen en resultados más confiables. Por otro lado, los datos del banco permiten establecer conexiones entre clientes y no clientes, por lo tanto, es posible dimensionar de forma directa el potencial impacto para campañas de apertura. Estos factores, sumado a las variables demográficas, de consumo y de vinculación con el banco permiten complementar el potencial de los modelos de redes a estudiar.

2.3 Hipótesis y posibles alternativas de solución

Como se mencionó anteriormente, el banco en la actualidad es la institución líder en el mercado en cuanto al número de tarjetas de crédito emitidas y por lo tanto la cartera de clientes del banco llega a una proporción importante de la población del país.

Esto, sumado al hecho de pertenecer a un holding con variadas unidades de negocio, provoca que se cuente con importantes fuentes de información tales como: variables demográficas de los clientes, variables que describen el nivel de vinculación a las distintas unidades de negocio o variables de propensión de reaccionar frente a ciertos ganchos o programas.

Bajo el enfoque que ahondan las publicaciones descritas anteriormente y considerando el potencial de las fuentes de información que cuenta el banco, se tienen las siguientes hipótesis para el desarrollo de este proyecto:

1) La base de transferencias bancarias permite construir redes de personas, clientes y no clientes del banco, sobre las cuales es factible calcular métricas que deriven de la estructura intrínseca de la red que conforman.

2) Una campaña de apertura, en particular una campaña de referidos, donde se ofrece un incentivo por cada persona que se “invita” a ser cliente del banco, permite contrastar la importancia de un cliente, por lo tanto, los clientes que son identificados como “más relevantes” tienen, en promedio, una tasa de respuesta superior al resto de los clientes.[5]

3) Las métricas obtenidas de las redes permiten determinar la importancia de un cliente en base a sus conexiones, pudiendo identificar a aquellos clientes que son más relevantes para el banco.

4) Una campaña de referidos permite determinar qué características del cliente son más relevantes para explicar la importancia de este. Diferenciando entre sus conexiones y métricas obtenidas del grafo.

2.4 Descripción del proyecto

[5] Con tasa de respuesta se entiende la cantidad de personas referidas por cliente que recibe tratamiento.

El proyecto se centra en la construcción de redes de clientes estudiado bajo los conceptos de teoría de grafos en conjunto con variables demográficas y variables que describen el niel de vinculación de los clientes con el banco. Con esto se busca discernir si es posible clasificar a aquellos clientes más relevantes para el banco respecto a campañas de apertura de tarjeta de crédito.

Una red (grafo) se entiende como una colección de nodos (vértices) conectados por aristas (enlaces). Se caracterizan por permitir estudiar interrelaciones e interacciones entre las unidades tanto a nivel individual como a nivel de comunidades. A partir de estas interrelaciones se pueden incorporar extensiones más generales o complejas de forma de conseguir métricas que describan, cuantifiquen o clasifiquen los nodos de una red.

En este proyecto, se busca construir redes sobre una base de clientes y no clientes del banco. Para esto, se cuenta con información que proviene de las transferencias realizadas por los clientes. Una transferencia bancaria se compone de un emisor, un receptor y un monto de dinero, contando con los elementos necesarios para ser utilizadas en la construcción de una red ya que en esencia establecen relaciones entre dos personas.

Para efectos de este proyecto, cada nodo (vértice) de la red representa una persona, la que se puede categorizar por cliente y no cliente. Las conexiones (enlaces) se establecen a partir de las transferencias entre dos personas. Los datos de transferencias disponibles corresponden a las emitidas o recibidas por algún cliente del banco.

Dado que las transferencias bancarias pueden tener múltiples ocurrencias en un periodo de tiempo, se aprovecha este hecho para caracterizar las relaciones. Esta caracterización se basa en un análisis *RFM* (*Recency Frequency Monetary* - explicado en el punto 4.3 del marco conceptual-) sobre las distintas transferencias realizadas entre dos personas.

Las variables de interés bajo las que se estudia la importancia de un **cliente** quedan definidas por dos dimensiones, **nodo** y **enlace**.

La relevancia bajo la dimensión **nodo**, se construye a partir de las métricas que se obtienen de la red.

La relevancia bajo la dimensión **enlace**, se construye como el promedio del “valor” de las relaciones de un cliente, siendo este valor obtenido mediante la aplicación de modelos de Machine Learning sobre las características obtenidas de una relación bajo el enfoque *RFM* que se mencionó anteriormente.

Resulta importante recalcar que en la dimensión **nodo** las métricas derivan de la forma intrínseca en que los nodos se ordenan dentro de una red, por lo tanto, no se toma en cuenta cual es el “valor” de esta relación entre dos personas. Por ejemplo, si dos personas se han transferido en múltiples oportunidades, para efectos de la dimensión **nodo** solo importa que la relación exista, mientras que para la dimensión **enlace**, lo importante son las características temporales de esta relación, es decir, cada cuanto tiempo ocurre, que montos de dinero se transfieren, etc.

El desarrollo del proyecto se puede dividir en 5 grandes etapas.

1. Selección de relaciones: En esta primera etapa se procesan los datos de la base de transferencias, obteniendo las métricas que van a describir cada una de las relaciones. A partir de esto se obtiene una tabla con relaciones únicas entre un nodo de origen, un nodo de destino y las variables que describen dicha relación.

2. División de base relaciones: En esta segunda etapa se seleccionan dos subconjuntos de la base procesada anteriormente. Un primer subconjunto contiene solo aquellas relaciones que derivan de transferencias emitidas entre enero de 2017 a febrero de 2018. El segundo subconjunto contiene aquellas relaciones que hayan ocurrido entre agosto de 2018 y agosto de 2019.

3. Entrenamiento de modelos: Se utiliza la base de relaciones con datos entre enero de 2017 a febrero de 2018, con el objetivo de entrenar un modelo de Machine Learning que permita determinar un score para cada relación. El modelo es entrenado usando data histórica de campañas de apertura implementadas durante abril de 2018. A partir de este score es posible calcular la dimensión **enlace** para cada nodo como el promedio del score de las relaciones que posee.

4. Modelamiento de redes: Se construyen dos redes distintas. La primera red se construye sobre el primer subconjunto de relaciones con datos hasta abril de 2018. Mientras que una segunda red es construida con el segundo subconjunto de relaciones con datos (agosto 2018 a agosto 2019).

5. Campaña piloto: Se diseña como piloto del proyecto una campaña de referidos. Con ella se busca contrastar los resultados obtenidos del análisis de relevancia de los clientes, permitiendo discernir si efectivamente aquellos conjuntos con clientes “más relevantes” tiene una mejor tasa de respuesta frente a una campaña de referidos que el resto de clientes “no relevantes”.

2.5 Propuesta de valor de las posibles soluciones

En este proyecto, se busca generar una herramienta que permita identificar a los clientes más relevantes para el banco en cuanto a su potencial de apertura. Bajo este contexto, los análisis respecto a los modelos aquí implementados van a mostrar si es posible o no discernir entre la “relevancia” de un cliente.

En caso de que los resultados sean positivos, se podrían aprovechar los insights generados para potenciar la gestión de las campañas de apertura contando con un perfilamiento de los clientes más relevantes, generando campañas que focalicen esfuerzos sobre aquellos clientes que se espera tenga una tasa de respuesta mayor. Por otro lado, se abre un espacio para nuevas investigaciones donde, entre otras cosas, sea posible evaluar la sensibilidad a los incentivos dado el nivel de relevancia del cliente.

Por otro lado, será posible establecer los lineamientos que deben seguir las investigaciones futuras para generar insight a partir de las herramientas de grafos hasta ahora no aprovechadas. Se observan dos grandes espacios de investigación donde se podría generar impacto, los programas de fidelización y la gestión de reclamos.

Los programas de fidelización son de gran relevancia para el banco debido a que ayudan en la captación de nuevos clientes y en la retención de los clientes más rentables. Esta situación también repercute en el resto de los negocios del holding. Actualmente, el principal programa de fidelización del banco es en base a puntos. Este programa entrega beneficios diferenciados tanto en la acumulación como en el canje de acuerdo con la categoría de un cliente. Dicha categoría depende directamente del nivel de consumo de un cliente. Bajo el enfoque de redes, no considerar el poder de influencia de un cliente sobre las personas con las que interactúa genera una clasificación errónea, por lo tanto, se podría estar sub o sobre estimando la categoría de un cliente, perdiendo valor en la utilización del programa de fidelización.

Por otro lado, la gestión de reclamos también resulta interesante de estudiar con un enfoque de redes. Bajo el supuesto que los clientes se muestran más abiertos a comunicar malas que buenas experiencias con las personas que se relacionan, la gestión de solicitudes de clientes de mayor relevancia (en base a sus conexiones) podría pasar por protocolos más completos y rigurosos. De esta forma, se podría evitar que los tramites o solicitudes se transformen en una mala experiencia o directamente en un reclamo debido a los costos que genera los comentarios y difusión negativa del cliente con las personas que se relaciona.

3 OBJETIVOS

3.1 Objetivo general:

Implementar una metodología de teoría de grafos y variables que describen el comportamiento de interacciones de clientes con otras personas, para identificar a aquellos de mayor relevancia en base a su poder de influencia, con foco en campañas de apertura de tarjeta de crédito.

3.2 Objetivos específicos:

- Generar las variables que permitan caracterizar una relación entre dos personas.
- Generar un modelo que permita cuantificar relaciones entre dos personas.
- Construir redes de clientes.
- Generar métricas a partir de redes que caractericen a los clientes.
- Diseñar e implementar una campaña piloto.
- Perfilar a los clientes analizados de acuerdo a su relevancia según las dimensiones calculadas.
- Generar recomendaciones a partir del conocimiento generado y oportunidades identificadas.

4 MARCO CONCEPTUAL

El proyecto será abordado mediante modelos de teoría de grafos y machine learning. Para esto se presenta un contexto de definiciones de los aspectos de ambas temáticas a utilizar.

4.1 Teoría de Grafos:

Teoría de grafos es una rama de las matemáticas y ciencias de la computación que estudia las propiedades de los grafos. Formalmente un grafo se constituye como una pareja ordenada de la forma $G = (V, E)$ donde V es el conjunto de vértices (Nodos) y E es el conjunto de aristas (Enlaces).

4.1.1 Propiedades de nodos

Dada la estructura de cómo se compone una red, existen distintas propiedades que caracterizan a cada nodo a partir de ciertas métricas llamadas medidas de centralidad. Si bien existe una extensa gama de propiedades y variaciones de ellas, las utilizadas en este proyecto se describen a continuación.

4.1.2 Eccentricity (Excentricidad)

La excentricidad de un nodo se define como la mayor distancia entre él con el nodo más alejado dentro de la misma red.

4.1.3 Centrality (Centralidad)

La centralidad se define como el número de conexiones que un nodo posee en una red.

4.1.4 Eigenvector Centrality (Centralidad vector propio)

La centralidad de vector propio define la relevancia de un nodo en base a la relevancia de los nodos con los que está conectado. Esta medida busca incorporar el hecho de que un nodo define su importancia de acuerdo al tipo de nodos con los que está en contacto y no tan solo del número de conexiones.

4.1.5 Betweenness (Intermediación)

La intermediación corresponde a la frecuencia con que un nodo ocurre en el punto geodésico (camino más corto) entre un par de nodos pertenecientes a la misma red. Esta medida permite cuantificar los nodos que juegan un papel importante en el flujo de comunicación de la información.

4.1.6 Nodos terminales

Se define como nodos terminales aquellos con el número más bajo de conexiones dentro de la red.

4.1.7 Sub grafo

Corresponde a un grafo que se construye sobre un subconjunto de nodos y enlaces de un grafo inicial.
Un subgrafo se utiliza para poder filtrar a los nodos aislados de los análisis, es decir, aquellos nodos o conjuntos de nodos que quedan desconectados de la red.

4.1.8 Filtro de redes

Se entiende por filtro de redes al proceso de eliminación de nodos de un grafo. Este proceso se suele aplicar sobre los subgrafos buscando reducir el tamaño de la red hasta un tamaño que permita el procesamiento de las métricas en un tiempo aceptable.

De las métricas mencionadas anteriormente tanto Betweenness como Eccentricity, son computacionalmente muy costosas de procesar, siendo el tiempo de procesamiento de ambas de crecimiento exponencial respecto al número de nodos y enlaces analizados en un grafo[6].

4.2 KPI

Un KPI (Key Performance Indicator) tal como su nombre lo indica corresponde a indicadores de desempeño. Se diseña para poder discernir entre distintas estrategias o análisis. En este caso en particular, se definirán KPI para determinar la calidad de un nodo, es decir que tan influenciador es sobre una red, KPI para determinar la calidad de los enlaces basados en una metodología RFM y KPI para determinar la calidad de un cliente basado en una combinación de los KPI de nodo y enlace.

4.3 Metodología RFM

Corresponde a una metodología descriptiva sobre el comportamiento transaccional de los clientes. En este caso, se utiliza para describir el comportamiento de las transferencias entre dos personas. La caracterización se logra en base a 4 métricas.

- Recency - ¿Hace cuánto fue la última transferencia entre estas personas?
- Frequency - ¿Qué tan frecuente se realizan estas transferencias?
- Monetary - ¿Cuánto dinero se transfieren?

[6] Frank W. Takes * and Walter A. Kosters, Computing the Eccentricity Distribution of Large Graphs.

4.4 Diseño experimental

El diseño experimental es una técnica que permite identificar las causas de un efecto estudiado. [7] En un experimento se crean situaciones previamente planificadas y diseñadas para generar una base de conocimiento.

4.4.1 Campaña piloto

Corresponde a la experimentación realizada sobre un subconjunto del total de datos. En el contexto de este proyecto, corresponde a la aplicación de una campaña de referidos sobre don conjuntos de clientes. El primero, corresponde a clientes analizados en la red, mientras que el segundo se obtiene de forma aleatoria sobre la base de clientes del banco. Esto se hace para obtener evidencia que permita analizar estadísticamente el efecto de las métricas que indican la relevancia de un cliente en contraste con su desempeño en las campañas de apertura.

4.4.2 Campaña de referidos

Campaña donde se ofrece un incentivo a los clientes (gancho) por invitar a personas a volverse clientes de una empresa. Para este proyecto, las campañas de referidos tienen un gancho en puntos que pueden ser canjeados por distintos productos o servicios tanto en el banco como en el resto de los negocios del mismo holding.

4.4.3 Tasa de lectura

Corresponde al porcentaje de clientes que abren el correo que comunica una campaña.

$$Tasa\ de\ lectura = \frac{Total\ correos\ abiertos}{Total\ correos\ enviados}$$

Ecuación 1: Tasa de lectura correos

4.4.4 Tasa de respuesta

Se define como el porcentaje de clientes que participan en una campaña respecto al total de clientes contactados. Aplicado a este proyecto, se entiende que un cliente participa en la campaña cuando está **refiriendo** al menos a una persona y que es contactado cuando recibe el mail con la campaña.

$$Tasa\ respuesta = \frac{Total\ clientes\ refieren}{Total\ clientes\ reciben\ campaña}$$

Ecuación 2: Tasa de respuesta correos

[7] Probabilidad y Estadística Montgomery & Runger

4.4.5 Grupo de control

En la literatura, se define como grupo de control al subconjunto de cliente que no recibe una intervención y se utiliza para contrastar los resultados con el grupo que recibe el tratamiento.

Resulta importante destacar que, para efectos de este proyecto, el grupo de control también recibe el tratamiento. Esto se hace debido a que no se espera que los clientes participen en una campaña de referidos si no reciben un incentivo por hacerlo. En consecuencia, no es posible contrastar con un subconjunto de datos que no haya recibido el tratamiento.
El grupo de control utilizado corresponde a un subconjunto de clientes del banco que fueron seleccionados de forma aleatoria para participar en la campaña de referidos. Por no pertenecer a la base desde donde se analizaron las redes, no es posible realizar ningún análisis en cuanto a las dimensiones utilizadas para definir la relevancia de un cliente (nodo y enlace). Con este grupo es posible contrastar si los esfuerzos destinados a la construcción y análisis de clientes más relevantes en base a redes generan el incremental deseado respecto al envio una campaña masiva.

4.4.6 Test Z de diferencia en proporciones

Test que permite determinar si dos proporciones de dos muestras de poblaciones independientes son diferentes. En el contexto de este proyecto, se contrastan las tasas de respuesta de los distintos conjuntos de clientes (segmentados según las métricas que se estén analizando).

El test se centra en la diferencia relativa entre las dos proporciones denominadas P_1 y P_2. Lo que es equivalente a formular las siguientes hipótesis nula y alternativa:

$$H_0: P_1 = P_2$$
$$H_1: P_1 \neq P_2$$

Ecuación 3: Hipótesis test de diferencia en proporciones

La estimación combinada de la proporción muestral P se calcula a partir del número de clientes que refiere en cada conjunto x_i y del número de clientes en cada conjunto n_i. Así, la proporción muestral se calcula como:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Ecuación 4: Valor p ecuación diferencia en proporciones

El valor de P se utiliza para calcular el estadístico de prueba Z_{prueba}, que luego es contrastado en la tabla de distribución normal.

$$Z_{prueba} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}}$$[8]

Ecuación 5: Valor z ecuación de diferencia en proporciones

4.4.7 Efecto incremental

4.5 Métodos de balanceo

Una base de datos se dice desbalanceada si contiene muchos más elementos de una categoría que de otras. Estas se utilizan para resolver problemas generados sobre el desempeño de los algoritmos de clasificación, que al tener pocos casos de una categoría no son capaces de predecir estos valores con éxito.

4.5.1 Undersampling

Este método consiste en utilizar un subconjunto aleatorio de la clase más concentrada de modo que se reduzca la diferencia de elementos entre ambas categorías.

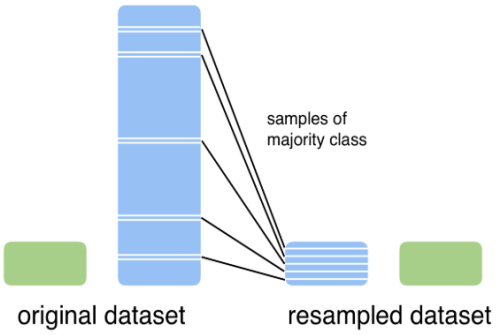


Ilustración 5: Método de balanceo Undersampling

4.5.2 Oversampling

Este método logra una base más balanceada duplicando elementos aleatorios de la clase con menos observaciones de modo que no se pierde información de la clase con más observaciones.

[8] towardsdatascience.com/everything-you-need-to-know-about-hypothesis-testing

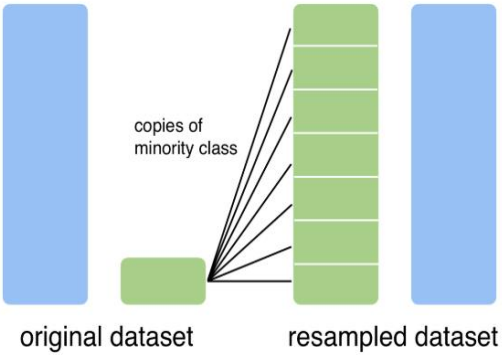


Ilustración 6: Método de balanceo Oversampling

4.5.3 Smote (Synthetic minority oversampling technique)

Este método fue diseñado para generar observaciones que sean coherentes con la distribución de la clase menos concentrada. Se basa en generar puntos intermedios "sintéticos" que se encuentran a una distancia intermedia de dos puntos de la clase desbalanceada. Este análisis se realiza en un espacio multidimensional, un ejemplo en dos dimensiones se puede observar a continuación:

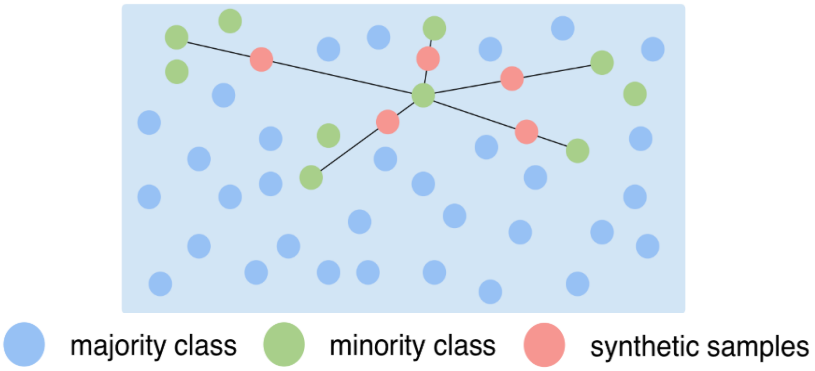


Ilustración 7: Método de balanceo SMOTE

4.6 Network Effect (Efecto red)

El Network Effect hace referencia a los potenciales beneficios sobre los clientes o usuarios de un producto/servicio cuando un nuevo cliente adquiere o utilice un producto/servicio. Un ejemplo de efecto red se observa en los medios de comunicación online. Un medio de comunicación con pocos usuarios tiene un bajo valor para cada uno por lo limitado que se vuelve la comunicación. Sin embargo, en caso de que aumente el número de usuarios, el medio se vuelve más valioso en cuanto permite a los usuarios establecer un mayor número de conexiones.

4.7 Algoritmos de aprendizaje supervisado

El aprendizaje supervisado corresponde a un tipo de algoritmos de machine Learning que se caracteriza por utilizar conocimientos previos en el entrenamiento de los modelos. Estos conocimientos son contrastados con los arrojados por los modelos para medir su desempeño. Por lo tanto, el objetivo del aprendizaje supervisado es aprender una función que, dada una muestra de datos y resultados deseados, se aproxime mejor a la relación entre los datos de entrada y salida.[9]

4.7.1 Linear regression (Regresión lineal)

Una regresión lineal es un tipo de análisis donde el valor de la variable dependiente puede ser estimado por medio de una combinación lineal de las variables independientes[10].

4.7.2 Logistic regression (Regresión logística)

A diferencia del modelo anterior, se utiliza para resolver problemas de clasificación donde se asume una relación lineal entre las variables independientes y la **transformación logit** de la variable dependiente. Dado esto, su resultado puede ser interpretado como la probabilidad de que un determinado evento ocurra[11].

Resulta importante destacar que los paquetes estadísticos entregan los mismos parámetros tanto para la regresión lineal como para la regresión logística. Sin embargo, la interpretación del efecto de una variable en la regresión logística se realiza sobre los Odds ratio y no sobre el coeficiente B_i como ocurre en la regresión lineal.

4.7.2.1 Odds Ratio

Se define como el ratio de la probabilidad de éxito sobre la probabilidad de fracaso. Por ejemplo, si un evento tiene una probabilidad de éxito de 0.8 (en efecto su probabilidad de no éxito es 0.2) su Odds ratio es de 4 a 1.

Para efectos de interpretación de los modelos de regresión logística, que un variable muestre un Odds ratio superior a 1 indica que tiene un efecto positivo sobre la probabilidad de éxito en la variable dependiente, mientras que si el Odds ratio es inferior a 1 entonces el efecto es negativo.

4.7.3 Decision tree (Arbol de decisión)

[9] https://towardsdatascience.com/supervised-vs-unsupervised-learning
[10] https://towardsdatascience.com/an-introduction-to-logistic-regression
[11] https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Un “decision tree” corresponde a una herramienta de clasificación que se basa en reglas lógicas que categorizan y particionan una serie de condiciones que ocurren de forma correlativa. [12]

4.7.4 Random forest

Random forest corresponde a la implementación de varios “decision tree” que operan de forma conjunta. Cada árbol predice una clase, la clase con más ocurrencias es la que arroja el modelo final. Dentro de los algoritmos de clasificación, corresponde a una herramienta robusta debido a que combina el poder predictivo de cada árbol los que son relativamente independientes (no correlacionados) entre sí.

4.7.5 Gradient boosting

Gradient boosting al igual que random forest corresponde a un algoritmo que se construye combinando una serie de decision tree. Se diferencia del random forest en que son ejecutados en paralelo. El algoritmo combina una serie de árboles donde cada uno es entrado para intentar corregir los errores del anterior. [13]

4.7.6 Ada boost

Ada boost corresponde a otro método iterativo de clasificación que se construye a partir de una combinación de múltiples clasificadores de bajo desempeño para obtener un clasificador de mayor robustez y precisión. Opera combinando los pesos de los clasificadores y entrenando el conjunto de datos de entrada en cada iteración lo que asegura la precisión de las predicciones.

5 METODOLOGÍA

5.1 Entendimiento del negocio (variables de interés)

Se busca analizar el contexto en el que se desarrollan las campañas de apertura en el banco, identificando áreas de trabajo involucradas, procesos relevantes y ordenes de magnitud de las estrategias y sus resultados. Con esta información interiorizada en el proyecto, se pretende generar nuevas estrategias para campañas de aperturas alineada y apalancada en los conocimientos que se desarrolle en el proyecto.

[12] https://towardsdatascience.com/understanding-random-forest-random-forest-random-forest
[13] https://towardsdatascience.com/comparing-different-classification-machine-learning-models-for-an-imbalanced-dataset

5.2 Entendimiento de los datos (Metodología KDD)

Debido a la gran cantidad de variables y fuentes de información el proyecto se abordará con una metodología KDD (*knowledge Discovery in Databases)* la que se describe a continuación:

- Identificación de objetivos
- Selección de variables
- Preprocesamiento
- Transformación de variables
- Data Mining y generación de modelos
- Interpretación y evaluación
- Integración al negocio

Este proceso es aplicado sobre las distintas fuentes de información que están siendo utilizadas para este proyecto, las que se describen a continuación:

- Base Datos Transferencia: Base que contiene los datos bancarios y de identificación tanto del emisor como del receptor de transferencias que se hayan realizado desde o hacia una cuenta vinculada al banco.

- Base Datos Pre Aprobados: Corresponde al filtro del área de riesgo sobre la base de datos de comportamiento financiero de todas las personas registradas en alguna entidad financiera y que es proporcionada por la CMF. Adicional a esto, existen personas que no se encuentran en la base de pre aprobados, sin embargo, tienen registro de comportamiento financiero y una edad sobre 21 años, estas personas aplican para una apertura de tarjeta.

- Base Datos Principalidad: Base que contiene el score de principalidad de cada cliente del banco. Este indicador se utiliza como una medida de la vinculación de los clientes con el banco y se calcula midiendo el uso en cuanto a productos y negocios del holding donde participa un cliente.

- Base Datos Clientes: Base que contiene información demográfica de los clientes o personas no clientes que hayan dado alguna vez el rut en caja en cualquier negocio del holding.

- Base Datos Registro Civil: Base que contiene las relaciones conyugales y de parentesco en primer grado para aquellas personas mayores de 18 años en Chile. A partir de estas relaciones se puede construir parentesco en segundo y tercer grado.

5.3 Diseño e implementación del piloto

Se diseña como piloto una campaña de referidos. La cual consiste en entregar incentivo en puntos a los clientes del banco por cada persona que estos refieren y que apertura una cuenta. Este tipo de campañas son gestionadas desde el área de tarjetas del banco desde donde se gestiona también el gancho o incentivo de la misma. Mediante esta

campaña se busca poder evaluar si efectivamente aquellas personas que fueron identificadas como clientes más relevantes para el banco generan una mayor cantidad de aperturas.

5.4 Evaluación de resultados del piloto

Una vez implementado el piloto, se debe evaluar los resultados obtenidos, esto implica contrastar la tasa de respuesta de la campaña de referidos tanto a nivel individual como grupal. Dado que, según el modelo planteado, la relevancia de los clientes se mide en dos dimensiones, se busca también cuantificar el valor porcentual del "score" propuesto para las dimensiones nodo y enlace, las que son calculadas a partir de la red (dimensión nodos) y a partir de los modelos de machine learning entrenados con datos de la campaña de referidos de 2018 (dimensión enlace). Los análisis sobre estas dos dimensiones se realizan con modelos de regresión logística que permiten evaluar el efecto de las métricas utilizadas sobre la tasa de respuesta de la campaña a nivel de cliente. Así, es posible discernir si el efecto observado es estadísticamente significativo. Además, se realiza una evaluación en modelos no supervisados donde se obtiene una categorización de los clientes en base a las mismas métricas. Esta categorización permite contrastar diferencias de las tasas de respuesta de cada grupo con las características de cada uno.

5.5 Planteamiento de investigaciones futuras

Finalmente, resulta importante ser capaz de materializar los resultados del proyecto en planteamientos concretos de cómo abordar las futuras investigaciones del tema.

Si bien el proyecto busca estudiar si es posible determinar la importancia de clientes contrastando con campañas de apertura, en el futuro se podrían contrastar con otras variables para estudiar si este efecto se repite también en otros contextos como modelos de predicción de fuga de clientes, utilización de alianzas o utilización de sucursales.

6 FLUJO METODOLOGICO

De la metodología anterior se desprende el macroproceso del proyecto el cual se compone de 5 etapas. Comienza por la selección de las bases de datos que definen las relaciones, una segunda etapa de procesamiento y segmentación de datos, una tercera etapa de entrenamiento de modelos de machine Learning, una cuarta etapa de modelamiento de redes y una quinta etapa de planificación e implementación del piloto para su posterior análisis y discusión de resultados. El desarrollo de cada etapa se detalla a continuación:

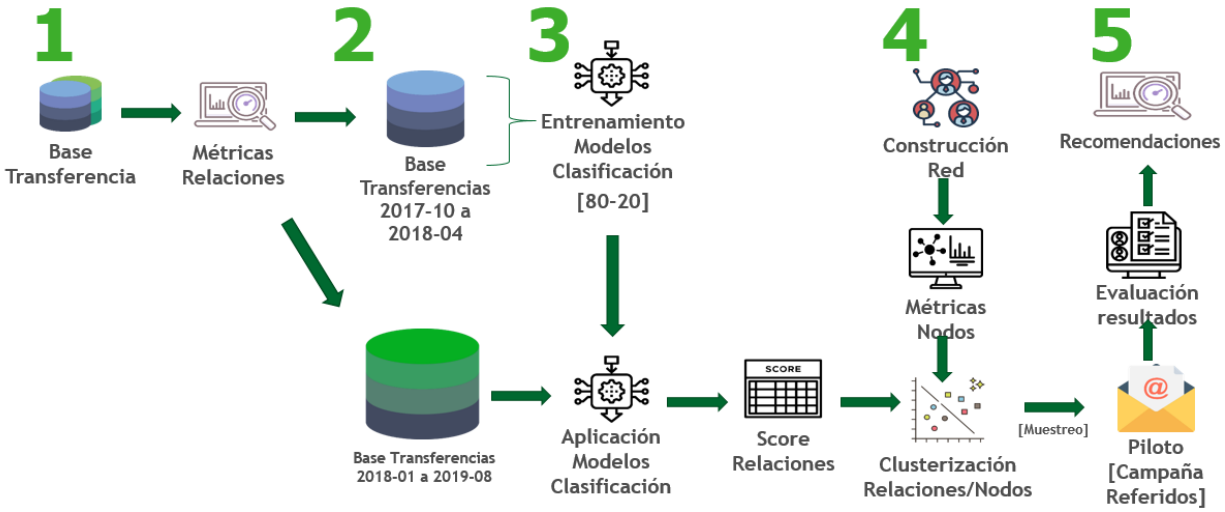


Ilustración 8: Macroproceso del proyecto

Etapa 1: Se utiliza la base de transferencia para establecer relaciones entre las personas. Dos personas que se hayan realizado al menos una transferencia desde enero de 2017 en adelante figuran como personas relacionadas. Cabe destacar que dado que la base contiene datos de transferencias que hayan sido emitidas o recibidas por un cliente del banco, las relaciones que aquí se analizan siempre comprometen a un cliente con un tercero.

Luego, sobre esta base se calculan métricas que caractericen la relación de transferencias entre dos personas. Se comienza por ordenar las tablas en rut origen y rut destino, siendo el rut origen el número mayor de la relación (recordando que para efectos de este análisis no importa la dirección del dinero de la transferencia).
Aplicando un análisis RFM sobre cada **conjunto** de transferencias entre dos personas se obtienen las métricas que describen su relación. Así, el resultado corresponde a una nueva base donde cada fila representa una relación **única** entre dos personas, tal como se puede apreciar en la siguiente ilustración:

- Tabla de entrada:

Rut origen	Rut destino	Monto [$]	Fecha
R_1	R_2	M_1	F_1
R_1	R_2	M_2	F_2
R_1	R_2	M_3	F_3

Ilustración 9: Ejemplo tabla transferencia

- Tabla de salida:

Rut origen	Rut destino	Monetary [$]	Frequency [Días]	Recency [Días]	RF
R_1	R_2	$\frac{(M_1 + M_2 + M_3)}{3}$	$\frac{F_3 - F_1}{2}$	$F_{hoy} - F_3$	$\frac{Recency}{Frequency}$

Ilustración 10: Ejemplo tabla de relaciones

Etapa 2: A partir de la tabla de relaciones se divide en dos bases independientes. Una primera base con transferencias desde enero de 2017 hasta febrero de 2018 y una segunda base con datos desde agosto de 2018 a agosto de 2019.

Luego, se filtra la base eliminando outliers respecto a cada métrica calculada para evitar observar efectos no deseados en los análisis, por ejemplo, por personas que utilizan su cuenta con fines comerciales, lo que para efectos del modelo podrían ponderar como importantes pero que no necesariamente derivan en el tipo de interacciones que se busca distinguir.

Finalmente, se utiliza una tercera base de datos que contiene los resultados de una campaña de referidos realizada durante abril de 2018. Esta base contiene el rut del cliente y de la persona referida. Luego, se ordena con el mismo criterio que la base anterior, es decir, se establece rut origen y rut destino según el número de rut mayor y menor respectivamente para luego cruzar con tabla de relaciones con datos del 2017.

De esta forma la tabla con relaciones únicas del año 2017 cuenta con una nueva variable llamada “Refiere”.

$$Refiere_{i,j} = \begin{cases} 1\ Si\ Rut\ i\ refiere\ a\ Rut\ j\ en\ campaña\ de\ referidos\ 2018 \\ 0\ No\ refiere \end{cases}$$

Ecuación 6: Definición de variable dependiente Refiere

- Tabla resultado cruce con campaña de referidos:

Rut origen	Rut destino	Monetary [$]	Frequency [Días]	Recency [Días]	RF	Refiere
R_1	R_2	…	….	…	…	0
R_2	R_3	…	….	…	…	1
R_2	R_4	…	….	…	…	1

Ilustración 11: Ejemplo tabla relaciones únicas con variable dependiente

Etapa 3: Utilizando el subconjunto de **relaciones** de 2017 se entrenó un modelo de clasificación utilizando como variable dependiente la variable "*Refiere*". A partir de este modelo es posible obtener la probabilidad de que una relación de un cliente con un no cliente se transforme en un referido en una campaña de apertura referidos valga la redundancia.

Luego se utilizó el segundo conjunto de **relaciones** con información desde agosto de 2018 a agosto de 2019 para correr el modelo entrenado anteriormente.

De esta forma, para cada relación de la base con datos de 2018/2019 se tiene la probabilidad que esa relación se transforme en un referido. Esta probabilidad es utilizada como "score" que representa el valor de una relación.

El proceso descrito queda representado gráficamente por la siguiente ilustración:

1. Entrenamiento modelo machine Learning:

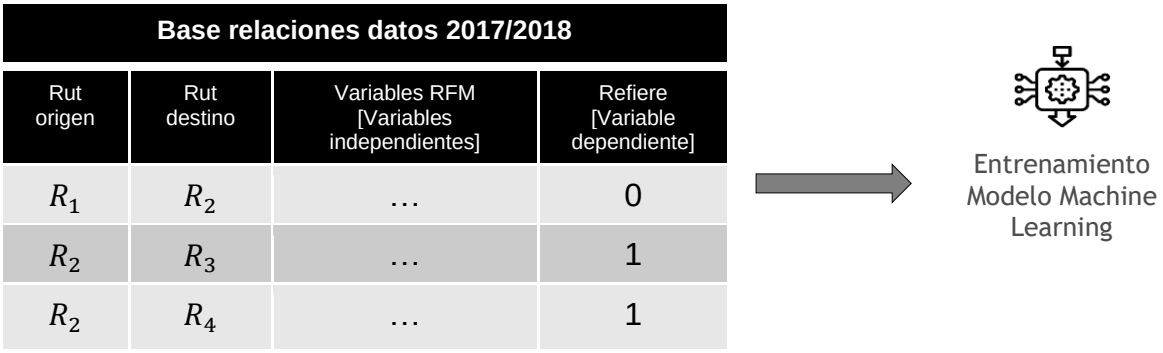


Ilustración 12: Ejemplo entrenamiento modelo machine learning

2. Clasificación modelo entrenado:

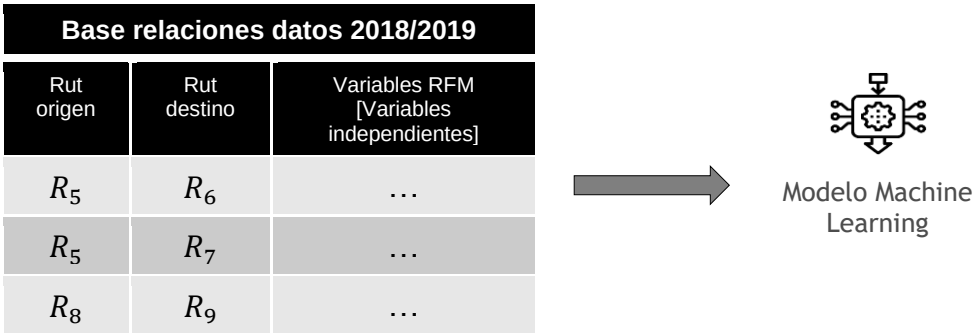
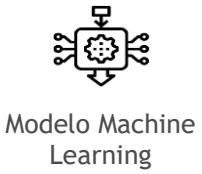


Ilustración 13: Ejemplo clasificación modelo machine learning

3. Resultado modelo machine learning clasificación modelo entrenado:



Nueva base relaciones [Datos 2018/2019]		
Rut origen	Rut destino	Score
R_5	R_6	n_{i}
R_5	R_7	n_{k}
R_8	R_9	n_{n}

Ilustración 14: Ejemplo resultado modelo machine learning

Etapa 4: Se utilizan las bases de relaciones para construir redes de clientes.

Una primera red se construye sobre las relaciones de enero 2017 a febrero 2018. Se calculan las métricas de **nodo** y se estudia una relación con los resultados de la campaña de referidos de 2018. A partir de este análisis es posible obtener una primera aproximación para entender el poder predictivo de estas métricas. Por otro lado, se busca aprovechar que la campaña fue enviada al total de clientes del banco, lo que podría ser un plus en la calidad de las métricas obtenidas.

Una segunda red se construye con los datos de relaciones desde agosto de 2018 a agosto de 2019. Se calculan las métricas derivadas de grafos y se promedian para construir la dimensión **nodo** de cada cliente.

Recordando que lo que se busca es poder determinar la relevancia de un cliente en las dimensiones, **nodo** y **enlace**.

Se calcula la dimensión enlace a partir del “score” entregado por el modelo de machine learning descrito en el punto anterior. Así, a partir de la tabla de relaciones entregada por el modelo, se promedia el score de las relaciones de cada cliente y se utiliza como valor de **enlace**.

Resulta importante mencionar que el cálculo de las métricas de redes requiere de un gran poder computacional por lo que el procesamiento se lleva a cabo en los servidores de google cloud platform (GCP). Además, las redes son filtradas mediante un algoritmo que, de forma iterativa, elimina los nodos terminales de la red hasta llegar a un grafo con 200.000 clientes o superior.

Etapa 5: La última etapa corresponde a la implementación de una campaña de referidos como piloto del proyecto. Con esto, es posible tener una tasa de respuesta sobre la cual se contrastan las métricas que describen la relevancia de un cliente bajo las dimensiones **nodo** y **enlace**.

Sé realiza un primer análisis estadístico donde, en conjunto con variables demográficas, se estudia el poder predictivo, bondad de ajuste y significancia estadística de las métricas calculadas.

Finalmente se realiza un análisis mediante modelos de clúster (no supervisado). Esto permite estudiar el efecto a nivel agregado, contrastando las diferencias en las tasas de respuesta de los distintos clústers y sus características.

7 ALCANCES

En este proyecto se busca establecer si es posible cuantificar la relevancia de un cliente con métricas derivadas de una red en conjunto con variables demográficas y de vinculación al banco.

Los alcances y limitaciones del proyecto se hacen presentes en distintos aspectos de la metodología, los que se abordan a continuación:

- Relaciones: Las relaciones se generan solo a partir de la matriz de transferencias, ignorando otras fuentes de información que también permiten establecer una relación entre dos personas, tales como, relaciones familiares o la relación entre clientes y sus adicionales de tarjeta de crédito. Estos aspectos no fueron considerados debido a que las transferencias permiten cuantificar una relación aplicando una metodología RFM lo que no es posible para relaciones familiares.

- Clientes: En línea con lo anterior, dado que las relaciones provienen de transferencias, se tiene que, de los clientes analizados en las redes, un 51% posee cuenta corriente y tarjeta de crédito, mientras que solo un 15% corresponde a clientes que solo poseen tarjeta de crédito. Esta proporción no es consecuente con el universo total de clientes del banco, donde solo un 18% posee cuenta corriente y tarjeta de crédito, mientras que un 66% corresponde a clientes que solo tienen tarjeta de crédito.

  Además, se encuentra el hecho de que las redes analizadas debieron ser filtradas de forma iterativa. Si bien este procedimiento se realiza tomando en consideración el efecto sobre la variable dependiente estudiada, se desconoce el impacto que podría generar sobre la calidad de las métricas de redes el hecho de analizar todas las relaciones.

- Variables independientes: Dentro del amplio espectro de variables demográficas y de vinculación al banco se utiliza solo un subconjunto menor de estas, excluyendo información que podría estar ayudando a explicar el efecto observado como el tipo de tarjeta que posee el cliente, datos de consumo y nivel de digitalización (esto considerando que la campaña invita a acceder a un formulario online). Otro aspecto a mencionar de las variables tanto demográficas como de vinculación, es que no es posible utilizarlas para los análisis de la campaña de referidos del 2018. Esto ya que para algunas de ellas la cantidad de missing values ronda el 50% de la base.

- Variable dependiente: Para ambas campañas la variable dependiente corresponde a “refiere”, tal como se mencionó anteriormente, hace referencia a si el cliente (o la relación) refiere a la otra persona. Esta variable se limita a los datos que ingresan al formulario de referidos y no está considerando que la persona referida efectivamente haya concretado la apertura pasando a ser cliente.

- Modelos: Los modelos de machine learning se entrenan con la variable dependiente refiere. En efecto, estos modelos entregan la probabilidad de que un cliente refiera a un no cliente, sin embargo, los modelos se utilizan para la clasificación de relaciones que se dan entre dos clientes. Este hecho se considera relevante pero no critico a la hora de los análisis. En primer lugar, debido a que de la base de transferencias se observa que la mayoría de las relaciones se da entre un cliente y un no cliente, y segundo, debido a que lo que se busca con los modelos es dar un “score” a cada relación pasando a ser menos relevante la probabilidad que realmente representa ese resultado.

- Gancho campaña: Para esta campaña se utilizó un gancho de 3.000 puntos por apertura tal como ocurre en la mayoría de las campañas de referidos. Con los análisis de este proyecto no es posible determinar si los resultados serían consistentes para campañas futuras con una variación del monto del gancho.

8 RESULTADOS ESPERADOS

Los resultados esperados del trabajo de título van de la mano con los objetivos planteados, los que se definen a continuación:

- Definición de la metodología para la creación de un análisis de redes sobre la base de clientes del banco.

- Set de relaciones que permita construir una red de clientes, considerando los criterios de filtro de la red que logren reducir el tiempo de procesamiento de las métricas minimizando la perdida de información.

- Set de modelos ya entrenados que determinen la relevancia de las relaciones analizadas.

- Modelos de aprendizaje no supervisado que permitan clasificar a los clientes, distinguiendo a los más relevantes.

- Propuestas de mejoras e investigaciones futuras.

9 DESARROLLO DE LA METODOLOGIA

Para cumplir con los objetivos e hipótesis del proyecto, el desarrollo metodológico se divide en 2 secciones.

En primer lugar, se aborda la campaña de referidos del 2018. En este análisis se busca entrenar los modelos que serán aplicados sobre la base de relaciones del 2019 y encontrar evidencia del efecto de las métricas de la dimensión **nodo** sobre los resultados de esta campaña.

La segunda parte del análisis se realiza sobre la campaña de referidos de este proyecto (campaña piloto). Allí se busca contrastar la relevancia de los clientes en las dos dimensiones propuestas, **nodo** y **enlace** con los resultados de la campaña**.** Respondiendo a sí es posible establecer la relevancia de un cliente en base a estas dos dimensiones. Esto se lleva a cabo tanto a nivel de cliente como a nivel agrupado.

9.1 Desarrollo metodológico 2018

Se presenta el desarrollo de la metodología para el primer subconjunto de datos que comprende datos de transferencias desde enero de 2017 hasta febrero de 2018.

9.1.1 Construcción base datos transferencias

Se comienza por seleccionar un subconjunto de la base de transferencias. Esta tabla contiene información que permiten caracterizar tanto al emisor como al receptor con los siguientes datos: Nombre, Rut, Banco, Tipo de cuenta. A la vez se cuenta información de la transacción como: Monto, fecha emisión y fecha contable.

Se selecciona un subconjunto de esta base con transferencias que se hayan emitido desde el 02 de enero de 2017 hasta el 28 de febrero de 2019. De las variables disponibles se seleccionan fecha y monto para describir la transacción, mientras que para el emisor y receptor se selecciona rut y banco.

La tabla resultante contiene 22.278.918 **transferencias**, 1.729.129 personas **únicas** de las cuales 343.072 son clientes del banco y 1.386.057 no clientes.

Se caracteriza las relaciones **únicas** que existen en la base. Utilizando los datos del banco de origen y banco de destino se distinguen 2 tipos de relaciones, entre dos clientes del banco y entre un cliente con una persona externa. El detalle se presenta a continuación:

Tipo relación		Numero relaciones únicas
Cliente	Cliente	240.255
Cliente	No Cliente	5.705.627
Total		5.945.882

Tabla 1: Caracterización relaciones 2017/2018

A modo de análisis exploratorio de la base seleccionada se estudia la distribución del monto de las transferencias. El monto promedio de las transferencias es de $117.748, sin embargo, el 50% de las transferencias se realizan por montos de $35.000 o menos, es decir, existen pocas transacciones por un muy alto valor que eleva el promedio. El detalle de la variable y su distribución se muestra a continuación:

- Monto

	MONTO
count	22.278.918
mean	117.748
std	376.114
min	500
25%	13.237
50%	35.000
75%	100.000
max	344.754.627

Tabla 2: Descripción monto transferencias 2017

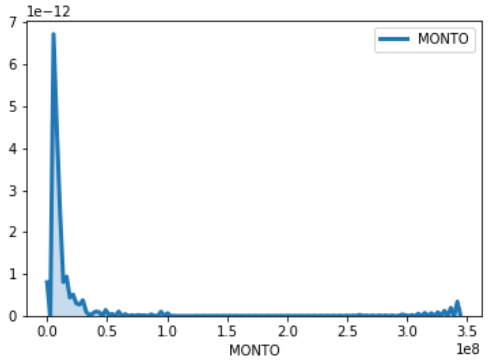


Ilustración 15: Distribución del monto de transferencias 2017

Mismo análisis de realiza sobre el número de interacciones entre **dos personas**, como se mencionó anteriormente, los 22.000.000 de transferencias equivalen a 5.945.882 de relaciones únicas. El promedio ronda las 4 interacciones, sin embargo, el 50% de las relaciones solo se da en 1 transferencia. El detalle de la variable y su distribución se muestra a continuación:

- Relaciones

	Relaciones únicas
count	5.945.882
mean	3,7
std	8,7
min	1
25%	1
50%	1
75%	3
max	3.257

Tabla 3: Descripción relaciones únicas 2017

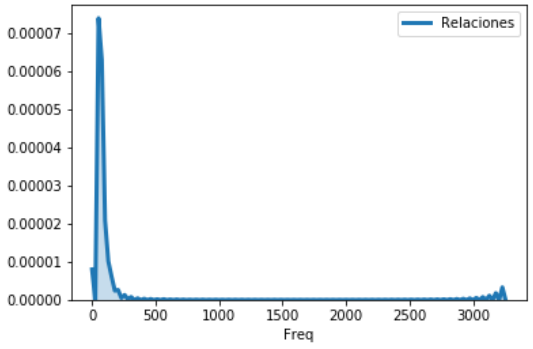


Ilustración 16: Distribución relaciones únicas de transferencias 2017

9.1.2 Construcción métricas relaciones

A partir de la base de transferencias se construye una de **relaciones**. Esta base se compone por relaciones únicas entre dos personas. Es decir, si en la tabla de transferencias dos personas se realizaban múltiples transferencias, cada transferencia quedaba representada en una fila. Para efectos de la base de relaciones, todas estas transferencias quedan en una sola fila con múltiples columnas que describen el comportamiento de las transferencias entre estas dos personas.

Las métricas generadas para cada relación son:

- Delta: Corresponde al tiempo en días desde la primera hasta la última transferencia.
- Frequency: Corresponde a la frecuencia medida en días con que dos personas se transfieren.
- Monto: Promedio del monto de las transferencias.
- Recency: Corresponde al tiempo en días que ha transcurrido desde la última transferencia y limite que se está considerando para la base.
- RF: Corresponde al ratio de Recency sobre Frequency. Representa los "ciclos normales" de interacción que han ocurrido desde la última transferencia[14].

9.1.2.1 Análisis exploratorio métricas

Dado que las métricas Frequency y RF no pueden ser calculadas para relaciones que solo se dan 1 vez, es decir, para dos personas que solo se han realizado una transferencia, se decide eliminar estos datos de la base. Esta acción se realiza ya que se busca representar las interacciones entre personas que perduran en el tiempo, donde una transferencia única no necesariamente es un indicio de aquello.

Así, la base resultante se compone de 2.305.606 relaciones. Sobre esta base se realiza un análisis exploratorio.

	DELTA	FREQUENCY	MONTO	RECENCY	RF
count	2.305.606	2.305.606	2.305.606	2.305.606	2.305.606
mean	167	47	101.983	109	12
std	134	54	263.088	106	51
min	1	0,033	500	1	0,002
25%	43	15	16.884	24	1
50%	136	30	39.825	71	2

[14] Por ejemplo: Si dos personas se transfieren con una frecuencia de 5 días y tienen un recency de 10 días, entonces RF será de 2. Es decir, han pasado dos ciclos de interacción en que estos clientes deberían haber interactuado y no lo han hecho.

75%	281	56	100.000	171	6
max	422	421	144.516.524	422	9.200

Tabla 4: Descriptivo tabla relaciones 2017/2018

A partir del descriptivo de las variables se observa que, a excepción de la variable RF, todas muestran un promedio muy superior a la media, es decir, hay indicios de una fuerte presencia de outliers.

Se estudia la distribución y Boxplot de cada métrica por separado para ver si es posible distinguir el mismo fenómeno. Los resultados se presentan a continuación:

1.1 Distribución variable Delta:	1.2 Boxplot variable Delta:

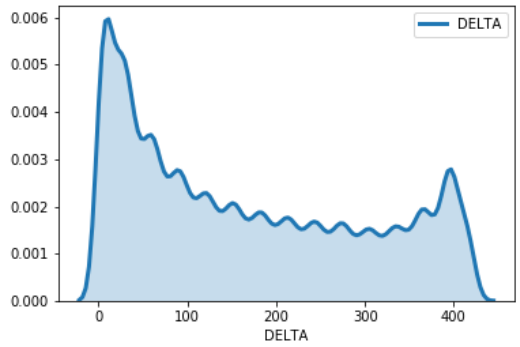
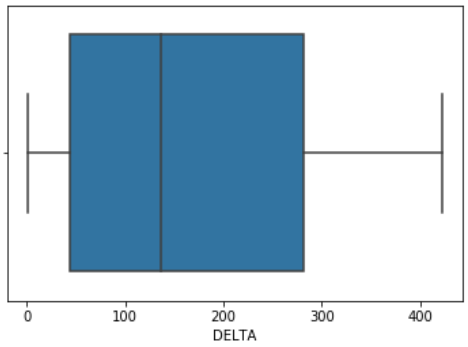

	

Ilustración 17: Distribución variable Delta	*Ilustración 18: Boxplot variable Delta*
2.1 Distribución variable Frequency:	2.2 Boxplot variable Frequency:

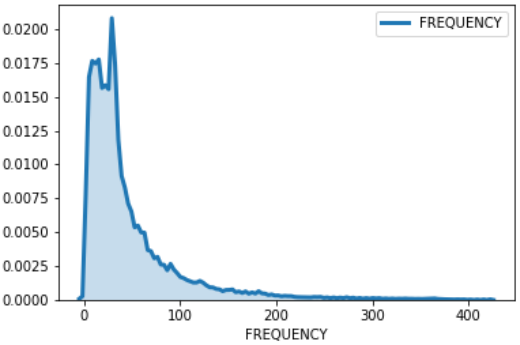
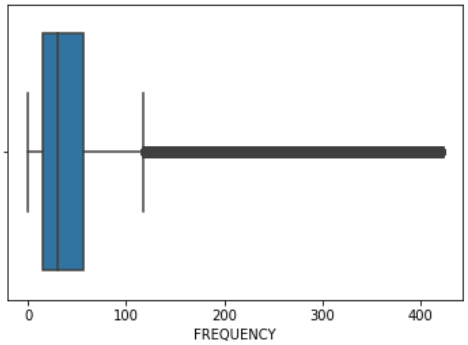

	

Ilustración 19: Distribución variable Frequency	*Ilustración 20: Boxplot variable Frequency*
3.1 Distribución variable Monto	3.2 Boxplot variable Monto

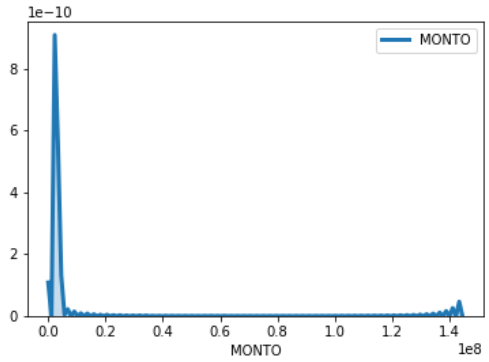


Ilustración 21: Distribución variable Monto

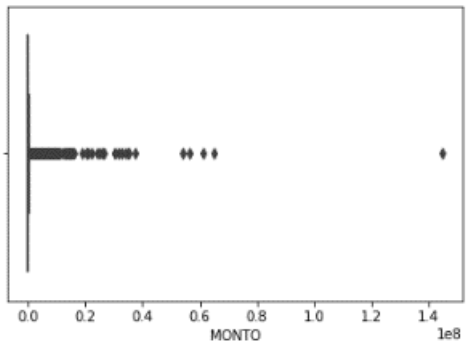


Ilustración 22: Boxplot variable Monto

4.1 Distribución variable Recency

4.2 Boxplot variable Recency

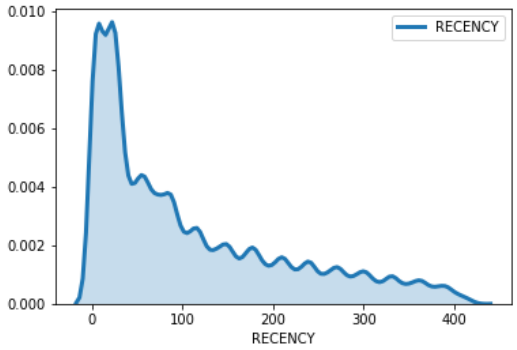


Ilustración 23: Distribución variable Recency

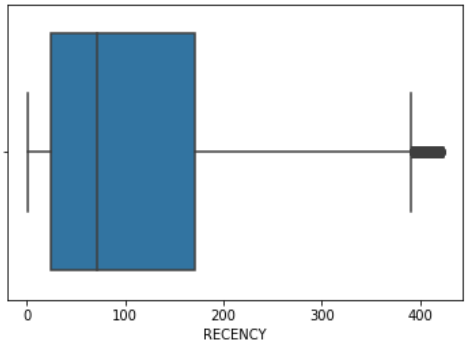


Ilustración 24: Boxplot variable Recency

5.1 Distribución variable RF:

5.2 Boxplot variable RF:

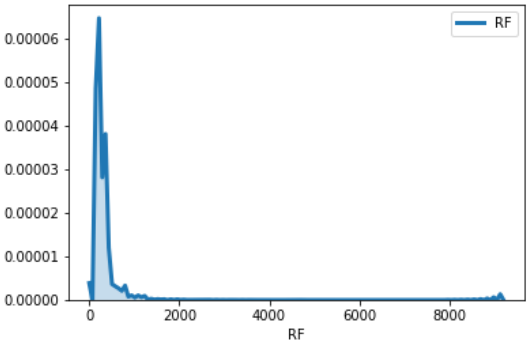


Ilustración 25: Distribución variable RF

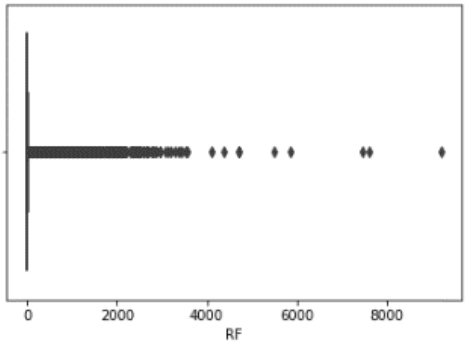


Ilustración 26: Boxplot variable RF

En las ilustraciones se observa que, tal como veía en los descriptivos, las variables tienden a estar concentradas en valores bajos. Siendo esta condición más severa para las variables Frequency, Monto y RF, las que muestran un numero preocupante de outliers.

9.1.3 Limpieza de datos

Para la limpieza de datos nulos o outliers se procesa la base de relaciones analizando cada variable por separado.

A partir de este análisis se filtran los outliers según el criterio de 3 desviaciones estándar. Es decir, para cada variable se calcula su promedio y desviación estándar. Luego, desde el promedio se considera un "tramo" con datos validos que va desde el valor promedio hasta 3 desviaciones estándar hacia arriba y hasta 3 desviaciones estándar hacia abajo. Los datos que quedan fuera de este tramo son eliminados. El detalle de criterio de filtro utilizado para cada variable se muestra a continuación.

Variable	Media	Desviación Estándar	Límite Inferior	Límite Superior	Datos Filtrados
DELTA	167	134	-234	569	27.363
FREQUENCY	47	54	-116	210	57.622
MONTO	101.983	263.088	-687.282	891.248	162.652
RECENCY	109	106	-210	427	30.218
RF	12	51	-142	166	325.305

Tabla 5: Filtro outliers base relaciones 2017-2018

Eliminando los outliers, la base resultante se compone de 1.702.446 relaciones.

9.1.4 Modelos machine learning

Los modelos de machine learning se utilizan para cuantificar, es decir, obtener un "score" de cada relación entre dos personas. Recordando que este "score" se utiliza para construir la dimensión **enlace** del estudio de la relevancia de los clientes basado en las dimensiones **nodo** y **enlace** antes planteado.

Siguiendo la estructura de la metodología, los datos de relaciones provenientes de enero 2017 a febrero de 2018 se utilizan para entrenar los modelos de machine learning.

Como se mencionó anteriormente, estos modelos se componen de un conjunto de variables independientes (en este caso las métricas de las relaciones) y una variable dependiente o variable a predecir.

La variable dependiente se construye a partir de una campaña de referidos realizada en abril de 2018.

9.1.4.1 Campaña referidos 2018

Esta corresponde a una campaña masiva, es decir, se envió a la totalidad de clientes contactables del banco.

El Funnel de contacto por email de la campaña, se muestra a continuación:

Envió campaña	Clientes contactados	Clientes refieren	Clientes aperturan
1.399.196 (100%)	352.614 (25,2%)	3.258 (0,9%)	796 (24,4%)

Tabla 6: Funnel campaña referidos 2018

De la tabla se observa que la tasa de lectura de la campaña fue de 25,2%, lo que va en línea con la tasa de lectura global del banco. La tasa de respuesta fue de un 0,9% lo que representa solo un 0,2% de los clientes contactados.

9.1.4.2 Variable dependiente

Recordando que la base analizada se compone de relaciones únicas, la variable dependiente se construye también en formato de relaciones a partir de la tabla de referidos de la campaña. Quedando definida con la siguiente estructura:

$$Refiere_{i,j} = \begin{cases} 1\ Si\ Rut\ i\ refiere\ a\ Rut\ j\ en\ campaña\ de\ referidos\ 2018 \\ 0\ No\ refiere \end{cases}$$

Ecuación 7: Definición de variable dependiente Refiere 2018

Considerando que solo un 0,2% de los clientes contactados por mail decide participar en la campaña, se decide estudiar la distribución de la variable dependiente.

En primer lugar, se realiza un nuevo filtro a la base de relaciones solo para efectos de entrenamiento de los modelos, eliminando las filas que contienen relaciones entre dos clientes. Esto se hace debido a que, por construcción, la variable dependiente no toma el valor 1 en esos casos (la campaña se envía a clientes para que refieran a no clientes). Con este nuevo filtro la base resultante se compone de 1.591.049 relaciones.

Se estudia la variable dependiente pudiendo notar que la base resultante queda fuertemente desbalanceada. La variable refiere toma el valor 1 solo en 383 casos. Es decir, del total de datos en la base de relaciones, solo 383 ocurrió que el cliente refirió a la otra persona. Tal como se puede apreciar en la siguiente ilustración:

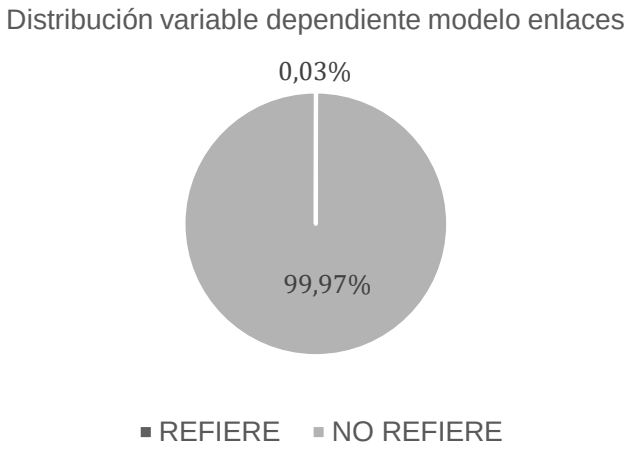


Ilustración 27: Distribución variable dependiente modelo enlaces

9.1.4.3 Métodos de balanceo

Se decide probar la implementación de 3 métodos de balanceo sobre la base de relaciones, Undersampling, Oversampling, y Smote.

Para decidir qué método de balanceo utilizar en los análisis posteriores, se corre una regresión logística permitiendo comparar las métricas de desempeño al utilizar los 3 métodos mencionados anteriormente. Los modelos se entrenan sobre un 80% de la base y son testeados con el 20% restante. A partir de este análisis se obtuvieron los siguientes resultados:

	Datos desbalanceados		Undersampling		Oversampling		SMOTE	
	Refiere	No Refiere	Refiere	No Refiere	Refiere	No Refiere	Refiere	No Refiere
Precision	0,00	1,00	0,79	0,71	0,00	0,96	0,00	1,00
Recall	0,69	1,00	0,72	0,78	0,00	1,00	0,71	0,73
F1-score	0,00	1,00	0,75	0,74	0,00	0,98	0,00	0,85
Accuracy	1,00		0,75		0,96		0,73	
AUC	0,74		0,77		0,84		0,77	

Tabla 7: Métricas de desempeño regresión logística para distintos métodos de balanceo

De la tabla anterior se observa que el método de balanceo Oversampling genera el modelo de mayor AUC, sin embargo, el modelo no es capaz de predecir el caso en que la variable dependiente toma el valor 1, por lo tanto, se descarta como opción de balanceo.

El método Undersampling parece dar al modelo métricas de desempeño en el rango adecuado, sin embargo, este método elimina datos de forma drástica quedando la base solo con un 0,04% de las filas, lo que podría generar pérdidas de información relevante.

Finalmente, el modelo SMOTE es capaz de predecir con un recall de 71% los casos positivos y logra un Accuracy y AUC de 73% y 77% respectivamente. Por lo tanto, se decide utilizar este método de balanceo por mostrarse un modelo equilibrado en cuanto a las métricas de desempeño.

9.1.5 Implementación modelos

Con el método de balanceo seleccionado, se prueban 4 modelos de clasificación: Regresión Logística, Random Forest, Gradient Boosting y Ada Boost. Estos se entrenan sobre un 80% de la base para ser testeados con el 20% restante. Se ejecutan sobre la base de datos normal y para la balanceada. Los resultados de cada uno se presentan a continuación:

	Logistic Regression	Logistic Regression [Balanced]	Random Forest	Random Forest [Balanced]	Gradient Boosting	Gradient Boosting [Balanced]	Ada Boost	Ada Boost [Balanced]
Precision	0,00	0,00	0,13	0,00	0,00	0,00	0,00	0,00
Recall	0,69	0,74	0,56	0,00	0,58	0,45	0,69	0,59
F1-score	0,00	0,00	0,21	0,00	0,00	0,00	0,00	0,00
Accuracy	0,74	0,60	1,00	1,00	0,91	0,80	0,76	0,65
AUC	0,73	0,72	0,49	0,54	0,60	0,62	0,70	0,63

Tabla 8: Métricas de desempeño modelos Machine Learning datos 2017/2018

De la tabla se observa que la regresión logística, tanto para los datos balanceados como no balanceados, posee el mayor AUC. Respecto al Accuracy, alcanza el valor máximo en el modelo Random Forest, pero considerando lo cercano a 0,5 del AUC, se desprende que el modelo está entregando una estimación similar a una asignación aleatoria. Así, juzgando por Accuracy el modelo Gradient Boosting es el que tiene el mejor desempeño.

Se debe mencionar que el Accuracy no es una métrica confiable para modelos desbalanceados por lo que se utiliza solo para comparar los modelos balanceados. El detalle de la curva ROC y auc de cada modelo se encuentra en los anexos (Anexo sección 9.1.5).

Finalmente, resulta importante recordar que los resultados de estos modelos se utilizaran para cuantificar el “score” de cada relación. Esto, pensando en la dimensión enlace de los clientes analizados en la base de datos que comprende relaciones de 2018/2019. Es decir, la base de relaciones que se está analizando se utiliza solo para entrenar los modelos.

En consecuencia, para efectos de este análisis solo se está estudiando la dimensión **nodo** de los clientes y por lo tanto las relaciones no tienen un valor, solo importa que la relación exista para plasmarlo en una red.

9.1.6 Modelamiento redes

EL modelamiento de redes tiene por objetivos construir la dimensión **nodo** de cada cliente. A partir de la base de relaciones resultante de la limpieza de datos se construye una red con clientes y no clientes. Resulta importante recordar que las redes son construidas solo en base a relaciones, es decir, no hay un score o relevancia.

Como se mencionó anteriormente, el cálculo de las métricas Betweenness y Eccentricity toman un tiempo de procesamiento exponencial en relación al número de nodos y enlaces analizados. En consecuencia, el volumen de datos de la base de relaciones requiere de varias semanas para el cálculo de estas métricas. Es por esto que se filtran las redes buscando acotar los tiempos de procesamiento a los plazos en que se enmarca el proyecto.

9.1.6.1 Filtro de red

Se realiza un primer filtro a la red seleccionando solo aquellos clientes que fueron contactados para la campaña de referidos y las personas con las que estos se conectan. Esto permite pasar de una red con 1.541.103 nodos y 1.702.446 enlaces a una de 1.153.059 nodos y 1.187.513 enlaces. Esta red se limita a los datos de clientes que se quiere estudiar, sin embargo, sigue siendo de un tamaño no procesable dentro de los plazos de este proyecto. Es por esto que se decide generar un método iterativo para filtrar los nodos terminales de la red, con esto se elimina a los nodos menos conectados generando una red más pequeña, pero que mantiene gran parte de las características iniciales ya que su estructura central, donde se concentran la mayor cantidad de conexiones, permanece inalterada.

En cada iteración se toma en consideración el número de clientes que participan en la campaña de referidos de la red resultante. No se debe olvidar que la relevancia de un cliente se contrasta con esta variable (refiere), por lo tanto, es importante mantener los casos positivos de la variable para no perder significancia estadística en los análisis posteriores.

El proceso iterativo se presenta a continuación:

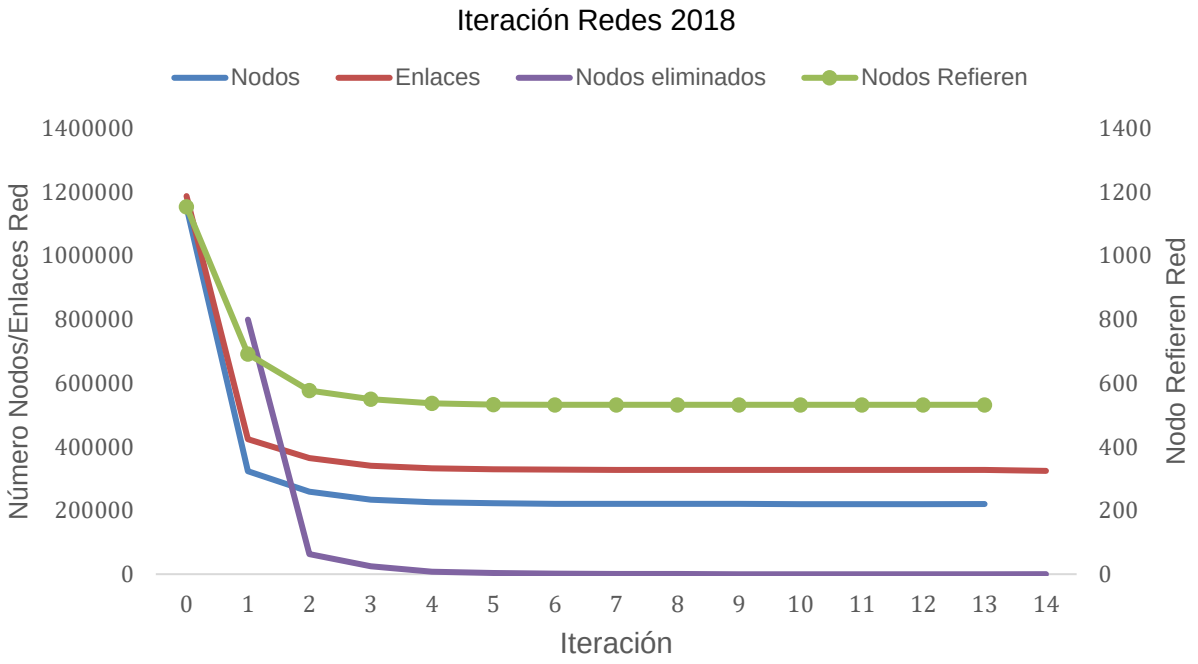


Ilustración 28: Iteración filtro de redes 2018

La red resultante se compone de 217.790 nodos y 324.434 enlaces. Respecto a la variable dependiente, se tiene que 515 clientes en la red participaron en la campaña de referidos. El detalle de cada iteración se encuentra en los anexos (Anexo sección 9.1.6.1)

9.1.7 Métricas redes

A partir de la red filtrada, se calculan las métricas que caracterizan a cada nodo. Luego, debido a que se busca explicar el comportamiento solo de los clientes, se realiza un filtro eliminando a los no clientes de los análisis, quedando un total de 125.527 nodos analizados.

En primer lugar, se estudia un descriptivo de las variables para entender el orden de magnitud y distribución de los datos, el detalle se observa a continuación:

	BETWEENNESS	CENTRALITY	ECCENTRICITY	EIGENVECTOR
count	217.790	217.790	217.790	217.790
mean	1.387.544	1,37E-05	28	3,9E-05
std	2.930.129	7,78E-06	2	2,1E-03
min	0	9,18E-06	22	5,4E-18
25%	242.306	9,18E-06	27	1,5E-09
50%	570.373	9,18E-06	28	1,1E-08
75%	1.370.816	1,38E-05	29	9,1E-08
max	127.000.000	9,18E-05	39	3,3E-01

Tabla 9: Descriptivo métricas red 2017/2018

De la tabla se observa que las variables están en ordenes de magnitud muy distintos. Se observa una gran concentración en valores muy pequeños para la variable eigenvector en contraste con Betweenness cuyos valores fluctúan en valores muy altos.

9.1.7.1 Análisis exploratorio métricas

Debido a los diversos ordenes de magnitud vistos en la tabla anterior, se decide normalizar las variables y estudiar su correlación con la variable dependiente refiere:

	BETWEENNESS	CENTRALITY	ECCENTRICITY	EIGENVECTOR	REFIERE
BETWEENNESS	100%	69%	-36%	6%	1%
CENTRALITY		100%	-33%	7%	2%
ECCENTRICITY			100%	-2%	-1%
EIGENVECTOR				100%	0%
REFIERE					100%

Tabla 10: Correlación métricas redes 2017/2018

A partir de la tabla se observa que existe una correlación importante entre las variables Centrality y Betweenness, al mismo tiempo que las correlaciones con la variable dependiente son bajas, pero van en el sentido que se esperara para cada una.
Resulta importante recordar que a excepción de la variable Eccentricity, todas las variables varían proporcional a la importancia de un nodo. Esto no ocurre con la variable Eccentricity, por lo tanto, se espera que su correlación con la variable dependiente sea negativa y positiva para las otras variables, tal como ocurre en este caso.

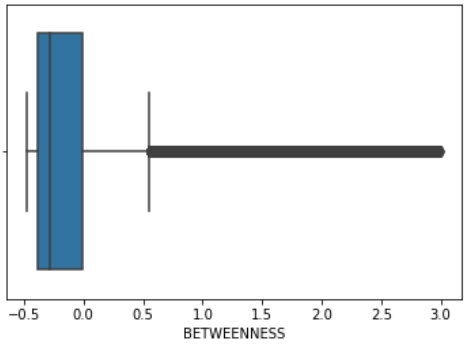
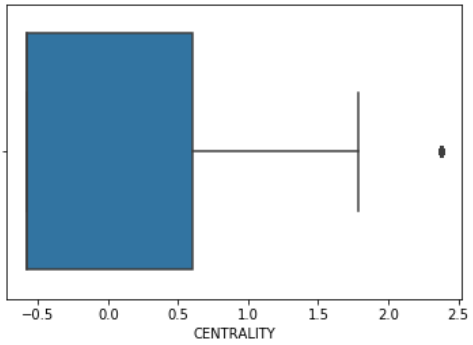
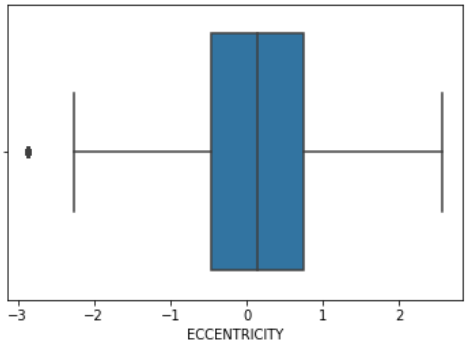
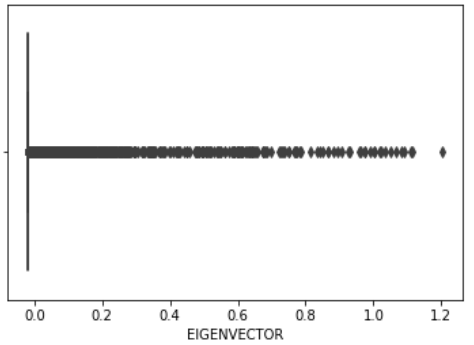
Debido a la alta dispersión que se observa en algunas variables, se realiza un filtro de outliers utilizando la misma técnica anterior, es decir, descartando datos se alejen más de 3 desviaciones estándar de la media.

Variable	Media	Desviación Estándar	Límite Inferior	Limite Superior	Datos Filtrados
BETWEENNESS	-1,31E-16	1,00	-3,00	3,00	3.535
CENTRALITY	-0,06	0,88	-2,69	2,58	4.431
ECCENTRICITY	0,04	0,99	-2,93	3,00	1.375
EIGENVECTOR	-0,01	0,41	-1,24	1,22	196

Tabla 11: Filtro outliers métricas red 2017/2018

A partir de la base sin outliers se vuelve a estudiar la correlación con la variable dependiente, observándose variaciones marginales en los resultados.

Se analizan los Boxplot de cada variable, tal como se puede apreciar a continuación:

1. Boxplot Betweenness	2. Boxplot Centrality
 *Ilustración 29: Boxplot Betweenness*	 *Ilustración 30: Boxplot Centrality*
3. Boxplot Eccentricity	**4. Boxplot Eigenvector**
 *Ilustración 31: Boxplot Eccentricity*	 *Ilustración 32: Boxplot Eigenvector*

A partir de la ilustración se puede apreciar que la variable Eigenvector, a pesar de estar normalizada, sigue concentrada en valores muy cercanos a cero. Efecto similar pero menos drástico se observa en la variable Betweenness.

Dado los bajos valores de correlación con la variable dependiente, se estudia la correlación ahora para transformaciones de cada variable, las transformaciones estudiadas son logaritmo natural, logaritmo base 10, transformación exponencial y cuadrática. Los resultados se pueden ver a continuación:

	CORRELACIÓN REFIERE
$\boldsymbol{ln(CENTRALITY)}$	2,22%
$\boldsymbol{log(CENTRALITY)}$	2,22%
CENTRALITY	2,15%
$\boldsymbol{exp(CENTALITY)}$	2,15%
$\textbf{CENTRALITY}^{\boldsymbol{2}}$	1,90%
$\boldsymbol{ln(EIGENVECTOR)}$	0,98%
$\boldsymbol{logCEIGENVECTOR)}$	0,98%
BETWEENNESS	0,87%

$\mathbf{BETWEENNESS^2}$	0,08%
$\mathbf{EIGENVECTOR^2}$	-0,06%
$\boldsymbol{exp(ECCENTRICITY)}$	-0,07%
$\boldsymbol{exp(EIGENVECTOR)}$	-0,10%
$\mathbf{EIGENVECTOR}$	-0,10%
$\boldsymbol{log(ECCENTRICITY)}$	-0,57%
$\boldsymbol{ln(ECCENTRICITY)}$	-0,57%
$\mathbf{ECCENTRICITY}$	-0,59%
$\mathbf{ECCENTRICITY^2}$	-0,61%

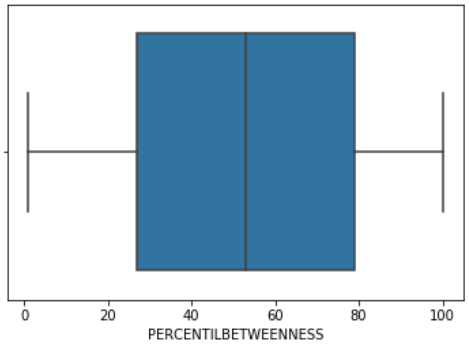
Tabla 12: Correlación transformaciones de variables

A partir de la tabla anterior se desprende que las transformaciones muestran una mejor correlación para todas las variables, sin embargo, el aumento en correlación es marginal por lo que se decide hacer caso omiso a este efecto.

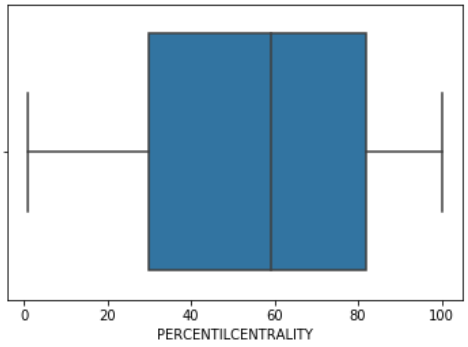
Finalmente, dada la gran dispersión de las variables en especial de eigenvector, se estudia cada variable en formato de percentiles.

En primer lugar, se estudia mediante Boxplot la distribución de las variables buscando el efecto de esta transformación sobre los outliers. Los gráficos se observan a continuación:

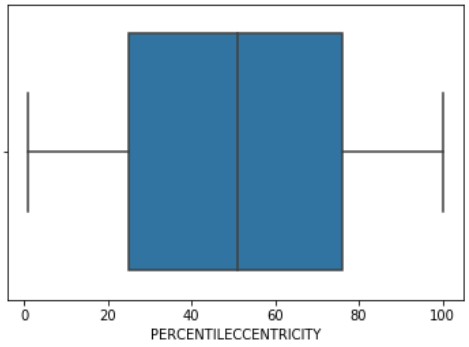
1. Boxplot Betweenness en percentiles



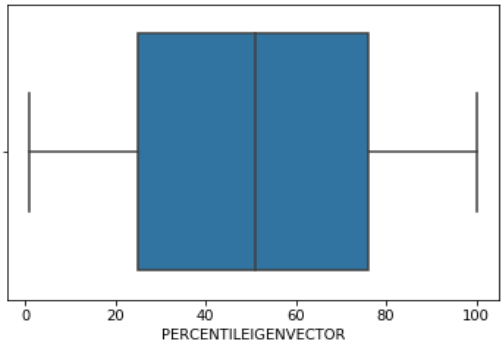
2. Boxplot Centrality en percentiles



3. Boxplot Eccentricity en percentiles



4. Boxplot Eigenvector en percentiles



Los Boxplot muestran un avance importante en la distribución de las variables, en especial lo que ocurre sobre Eigenvector y Eccentricity.

Buscando ahondar más en el efecto de esta transformación se estudia la correlación con la variable dependiente, el detalle se muestra a continuación:

	PERCENTIL BETWEENNESS	PERCENTIL CENTRALITY	PERCENTIL ECCENTRICITY	PERCENTIL EIGENVECTOR	REFIERE
PERCENTIL BETWEENNESS	100%	66%	-47%	50%	1%
PERCENTIL CENTRALITY		100%	-27%	31%	2%
PERCENTIL ECCENTRICITY			100%	-51%	-1%
PERCENTIL EIGENVECTOR				100%	1%
REFIERE					100%

Tabla 13: Correlación transformación percentiles

En las correlaciones no se observan variaciones significativas respecto a las variables normalizadas. Si bien los cambios son menores, se observa que las correlaciones con la variable dependiente siguen siendo consistentes con el signo esperado para cada variable, es decir, correlación negativa en Eccentricity y positiva en el resto de variables.

9.1.8 Análisis estadístico

Antes de entrar en los análisis de regresiones logística se debe recordar el concepto de "Odds ratio" explicado en el marco conceptual. Odds ratio se define como la probabilidad de acierto sobre la probabilidad de fracaso. El coeficiente que entrega la regresión logística corresponde a $\ln(Odd\ ratio_i)$ para la variable *" i "*. Es decir, para tener una interpretación del sentido del efecto se debe calcular el Odds como e^{B_i}.

En consecuencia, la interpretación del efecto de una variable se realiza mediante el valor de la columna Odds. En caso de que este valor sea cercano a 1 no es posible concluir un efecto sobre la variable independiente[15]. Para valores superior a 1 entonces la variable tiene un efecto positivo sobre la probabilidad de éxito. Análogo, pero en sentido contrario, es la interpretación para el caso menor que 1.

A continuación, se presentan las regresiones logísticas para los datos normalizados y en percentiles. Cabe destacar que se utilizó un 80% de la base para entrenar el modelo y un 20% para testarlo. Sobre este último porcentaje se construye la curva ROC.

[15] Dado que Odd ratio se define como $\frac{p}{(1-p)}$ el caso Odd ratio igual a 1 se produce cuando $p = 0.5$ en consecuencia no hay un efecto de esta variable sobre la variable independiente.

- Regresión logística datos normalizados

Variable	Coefficient	Odd	p-value
Intercept	-5,654 (***)	0,004	0,000
BETWEENNESS	-0,070	0,932	0,601
CENTRALITY	0,370 (***)	1,448	0,000
ECCENTRICITY	0,001	1,001	0,984
EIGENVECTOR	0,011	1,011	0,995

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Tabla 14: Regresión logística datos normalizados 2018

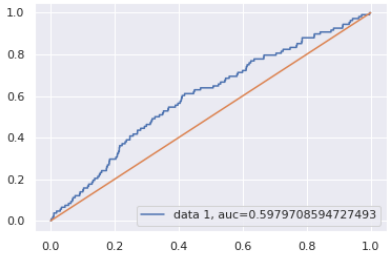


Ilustración 33:Curva ROC Regresión logística datos normalizados 2018

- Regresión logística datos percentiles

Variable	Coefficient	Odd	p-value
Intercept	-6,229 (***)	0,002	0,000
PERCENTIL BETWEENNESS	0,025	1,025	0,924
PERCENTIL CENTRALITY	0,889 (***)	2,430	0,000
PERCENTIL ECCENTRICITY	0,046	1,047	0,824
PERCENTIL EIGENVECTOR	0,234	1,264	0,263

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Tabla 15: Regresión logística datos percentiles 2018

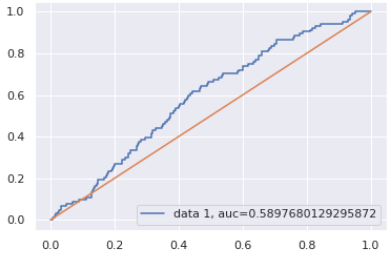


Ilustración 34: Curva ROC Regresión logística datos percentiles 2018

Respecto a la regresión con datos normalizados, se observa, que la variable que más afecta la variable dependiente es la centralidad, teniendo un Odds ratio cercano a 1,5 y significativa al 99,9%. El intercepto también es significativo al 99,9% con un Odds ratio muy cercano a cero. Lo que indica que el valor esperado de la variable dependiente tiende a ser 0.

Efecto similar se observa en la regresión con datos en percentiles, donde la variable centralidad es la más importante con un Odds ratio superior a 2 y significativa al 99,9%.

En ambos casos se observa una curva roc con un auc de 0,59, lo que da índicos que los modelos tienen un Accuracy similar.

Por lo que ya se mencionó respecto a la variable Eccentricity, se esperaría que tuviera un Odds ratio menor que 1 debido al efecto inversamente proporcional sobre la relevancia de un cliente. Esto no ocurre en ninguno de los dos modelos, lo que sumado al efecto del intercepto de cada modelo, da indicios de que no se cuenta con la bondad de ajuste suficiente para poder determinar la relevancia de las variables a nivel de clientes.

Por lo anterior, se decide realizar un análisis no supervisado, buscando evaluar los efectos a nivel agregado, es decir, en conjuntos de clientes.

9.1.9 Análisis No Supervisado

Se realiza un análisis no supervisado encontrando clúster con el algoritmo K Means. Este análisis se realiza primero para los datos normalizados y luego para la transformación en percentiles.

9.1.9.1 Análisis no supervisado variables normalizadas

Se comienza utilizando dos métodos para determinar el número de clúster a utilizar: Elbow (regla de codo) y Calinski Harabasz[16].

1. Elbow Method For Optimal k	2. Calinski Harabasz Method For Optimal k

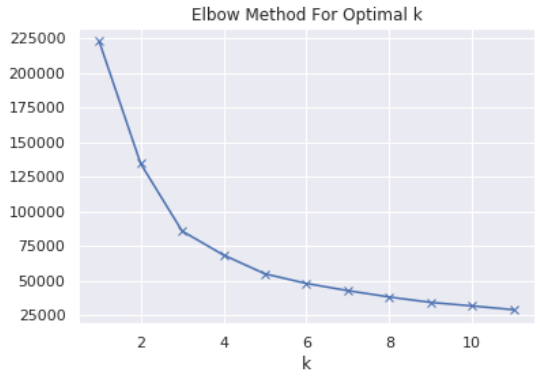
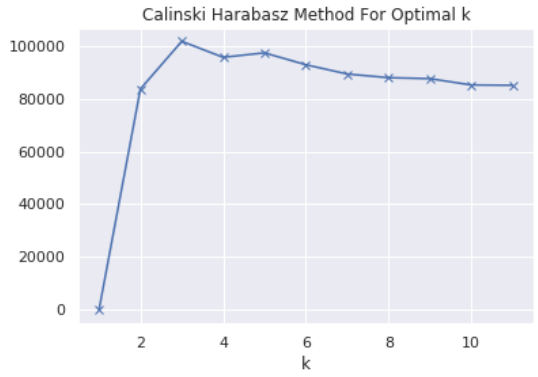

	

Ilustración 35: Elbow curve datos normalizados 2018	*Ilustración 36: Calinski Harabasz datos normalizados 2018*

A partir del grafico se decide utilizar 3 clústers. Con esto se da paso a una agrupación para estudiar: tasa de respuesta promedio de cada clúster; valor de las métricas de los centroides; número de clientes pertenecientes a cada clúster y una variable "Z" que representa el promedio de las métricas de los centroides. El número del clúster es asignado en orden descendente en relación a la tasa de respuesta del mismo, el detalle se puede observar a continuación:

Clúster	Tasa respuesta	BETWEENNESS	CENTRALITY	ECCENTRICITY	EIGENVECTOR	CLIENTES	Z
3	0,57%	0,6	1,28	-0,63	-0,01	24.729	1,23
2	0,34%	-0,22	-0,32	-0,4	-0,02	63.838	-0,95
1	0,30%	-0,33	-0,31	1,15	-0,02	38.960	0,48

Tabla 16: Agrupación clúster análisis no supervisado datos normalizados 2018

Se observa que para el mejor conjunto (clúster 3), la tasa de respuesta es cercana al doble del peor conjunto (clúster 1). A la inversa ocurre con el número de clientes, donde la proporción de clientes de este último es casi el doble que para el clúster 3.

[16] Para el motodo Elbow, el numero de cluster adecuado se encuentra donde se produce un cambio en la pendiente de la curva, mientras que para Calinski Harabasz se encuentra en el máximo de la curva.

Se realiza un test de diferencia en proporciones para determinar si la diferencia de las tasas de respuesta de los grupos es estadísticamente significativa.

Clúster	3	2	1
3	0	4.92 (99%)[17]	5.17 (99%)
2		0	0.93
1			0

Tabla 17: Test de diferencia en proporciones datos normalizados 2018

A partir de la tabla se tiene que el clúster 3 tiene una tasa de respuesta que es estadísticamente superior a la de los otros conjuntos. Sin embargo, la diferencia entre los clústeres 1 y 2 no es significativa, por lo tanto, no se puede concluir que sean conjuntos con un desempeño diferenciable.

Para ahondar más en las diferencias de los conjuntos se realiza un análisis de dispersion. Es importante recordar que, a excepción de Eccentricity, las métricas toman valores mayores en cuanto el cliente es más importante, por lo tanto, del análisis se espera que aquellos clientes más importantes, es decir, que pertenecen al clúster 3, estén concentrados en la esquina superior derecha. A excepción de cuando se estudia la variable Eccentricity donde debido al cambio de dirección de esta variable se deberían concentrar en el otro extremo.

1. Eigenvector-Betweenness	2. Centrality-Betweenness	3. Centrality-Eigenvector

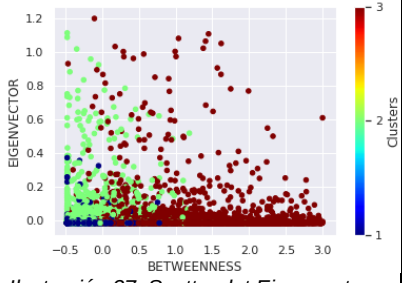
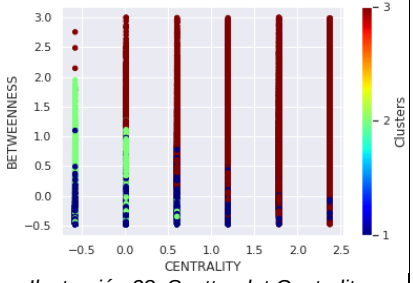
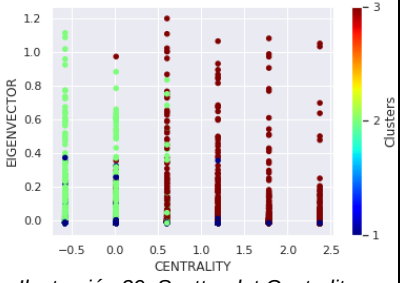

		

Ilustración 37: Scatterplot Eigenvector-Betweenness	*Ilustración 38: Scatterplot Centrality-Betweenness*	*Ilustración 39: Scatterplot Centrality-Eigenvector*
4. Eccentricity-Eigenvector	5. Eccentricity-Betweenness	6. Eccentricity-Centrality

[17] Nivel de significancia estadística

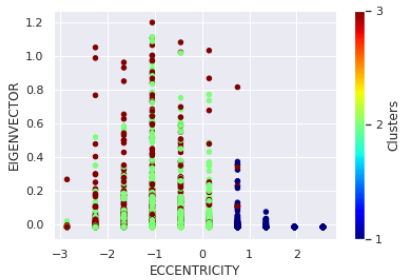


Ilustración 40: Scatterplot Eccentricity-Eigenvector

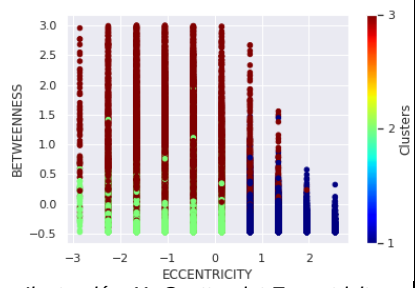


Ilustración 41: Scatterplot Eccentricity-Betweenness

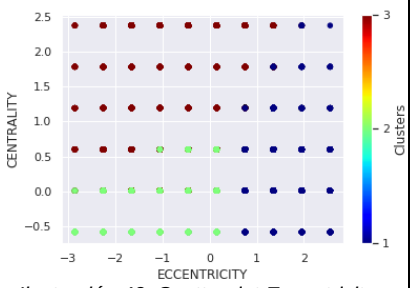


Ilustración 42: Scatterplot Eccentricity-Centrality

A partir del análisis grafico no es posible ver el efecto esperado o al menos no es claramente distinguible. Esto se atribuye a que algunas métricas (ver anexo sección 9.1.9) entregan como resultado valores enteros que provocan la concentración de los puntos, ensuciando el análisis gráfico.

Dado lo anterior, se decide realizar el mismo análisis, pero ahora en 3 dimensiones.

1. Eigenvector-Betweenness-Centrality	2. Betweenness-Centrality Eccentricity	4. Betweenness-Centrality-Eccentricity [solo clientes refieren]

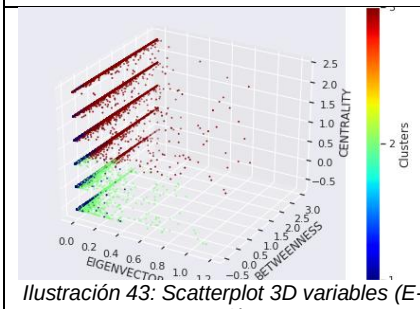
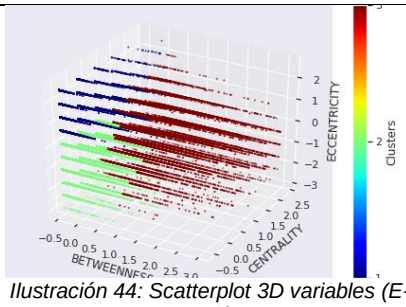
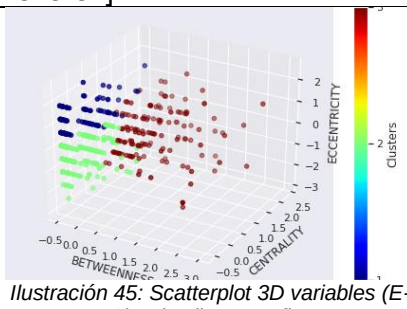

		

Ilustración 43: Scatterplot 3D variables (E-B-C)	*Ilustración 44: Scatterplot 3D variables (E-B-C)*	*Ilustración 45: Scatterplot 3D variables (E-B-C) solo clientes refieren*

A partir de la gráfica 1, se logra observar que la variable eigenvector, a pesar de estar normalizada, sigue con una concentración importante de valores cercanos a 0. En la gráfica 2, donde se realiza el mismo análisis excluyendo esta variable, se observa que, si bien los puntos se ven homogéneos dentro de cada clúster, no es posible distinguir que para el clúster 3 estos se posicionen en los extremos más altos de cada variable.

Por último, en la gráfica 3 se analiza solo a los clientes que refieren, donde tampoco es posible distinguir que los puntos del clúster 3 se distribuyan en el extremo esperado.

Dado que con los datos normalizados no es posible obtener conclusiones convincentes, se realiza el mismo análisis para las variables en percentiles.

9.1.9.2 Análisis no supervisado variables percentiles

Se comienza estudiando el número de clúster a utilizar con los mismos dos métodos anteriores, tal como se aprecia a continuación:

1. Elbow Method For Optimal k	2. Calinski Harabasz Method For Optimal k

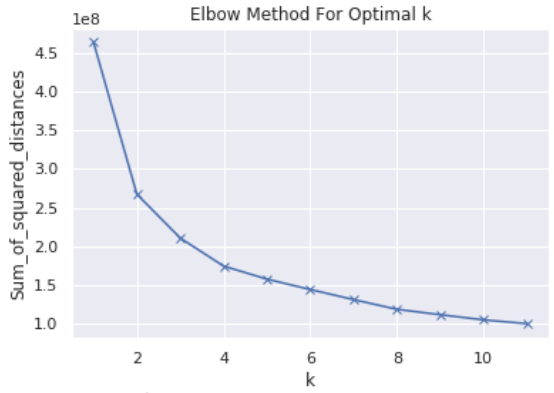
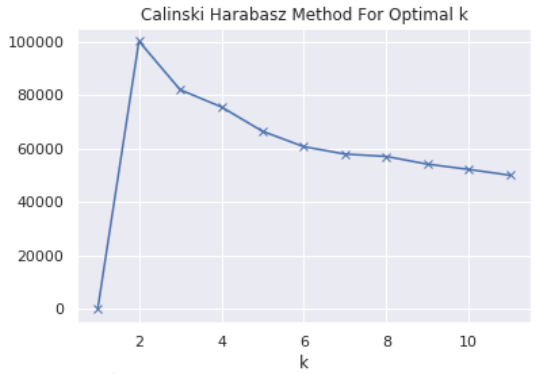

	

Ilustración 46: Elbow curve datos percentiles 2018	*Ilustración 47: Calinski Harabasz datos percentiles 2018*

Se observa que los modelos no entregan resultados consistentes entre sí, por lo tanto, se decide utilizar solo la curva Elbow utilizado 4 clusters.

Se realiza una agrupación calculando los mismos valores que la sección anterior, como se puede observar a continuación:

clúster	TASA RESPUESTA	PERCENTIL BETWEENNESS	PERCENTIL CENTRALITY	PERCENTIL ECCENTRICITY	PERCENTIL EIGENVECTOR	Clientes	Z
4	0,53%	84,75	83,99	25,38	75,93	38.264	270
3	0,46%	61,59	75,08	64,32	39,09	28.959	240
2	0,29%	22,59	31,95	78,77	21,19	35.703	155
1	0,26%	39,84	30,29	36,88	62,81	32.763	170

Tabla 18:Agrupación clúster análisis no supervisado datos percentiles 2018

De la tabla anterior se desprende que para el mejor conjunto (clúster 4), la tasa de respuesta es el doble que para el peor conjunto (clúster 1). Este número es consistente con la variación del valor promedio de las métricas de los centroides (variable Z) que, a excepción del clúster 2, se mueve en el mismo sentido que la tasa de respuesta de los conjuntos.

Respecto al número de clientes se observa que es relativamente equitativo entre los 4 conjuntos.

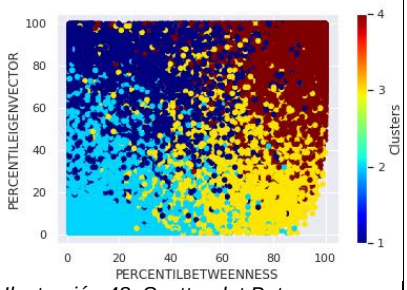
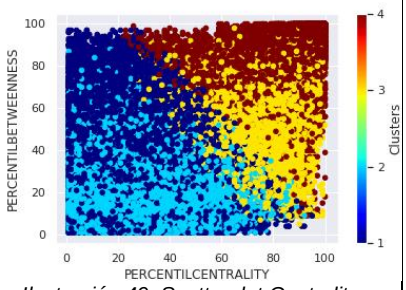
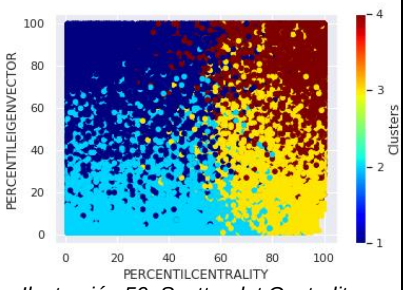
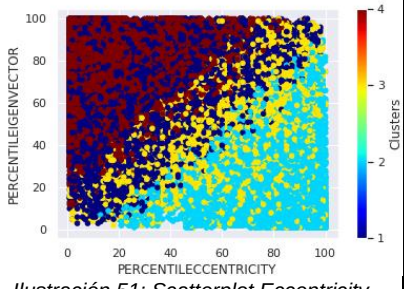
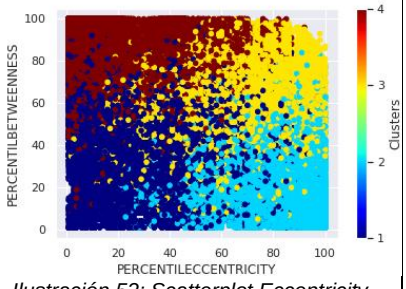
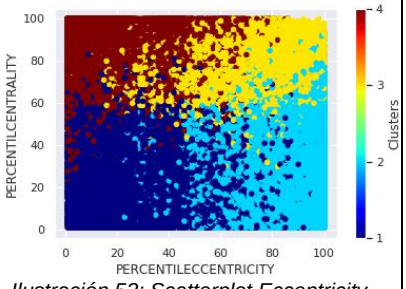
Se calcula un test de diferencia en proporciones para determinar si la diferencia de la tasa de respuesta de los grupos es estadísticamente significativa.

clúster	4	3	2	1
4	0	1.30 (80%)[18]	5.19 (99%)	5.74 (99%)
3		0	3.65 (99%)	4.25 (99%)
2			0	0.74
1				0

Tabla 19: Test de diferencia en proporciones datos percentiles 2018

A partir de la tabla se observa que los dos mejores conjuntos, clúster 3 y 4, tienen una tasa de respuesta estadísticamente mejor que los otros conjuntos. Respecto al clúster 1 y 2, la diferencia entre ellos no es significativa, por lo tanto, no se puede hablar de conjuntos diferenciables.

Al igual que en la sección anterior, se realiza un análisis de métricas conjuntas para estudiar la relación entre las combinaciones de variables para cada conjunto, el detalle se observa a continuación:

1.Betweenness-Eigenvector	2. Centrality-Betweenness	3. Centrality-Eigenvector
 *Ilustración 48: Scatterplot Betweenness-Eigenvector*	 *Ilustración 49: Scatterplot Centrality-Betweenness*	 *Ilustración 50: Scatterplot Centrality-Eigenvector*
4. Eccentricity-Eigenvector	**5. Eccentricity-Betweenness**	**6. Eccentricity-Centrality**
 *Ilustración 51: Scatterplot Eccentricity-Eigenvector*	 *Ilustración 52: Scatterplot Eccentricity-Betweenness*	 *Ilustración 53: Scatterplot Eccentricity-Centrality*

[18] Nivel de significancia estadística

A partir de las gráficas anteriores se logra distinguir con claridad la distribución en el espacio de los distintos conjuntos. Se aprecia claramente que los puntos pertenecientes al mejor conjunto (clúster 4) se concentran en la esquina superior derecha, a excepción de cuando se analiza la variable Eccentricity donde en este caso se ubican en la esquina superior izquierda lo que se encuentra en línea con la interpretación de esta variable.

Se realiza un análisis en 3 dimensiones para ver si los resultados son consistentes con los gráficos ya estudiados.

1. Eigenvector-Betweenness-Centrality	1. Eigenvector-Betweenness-Centrality [Solo clientes refieren]

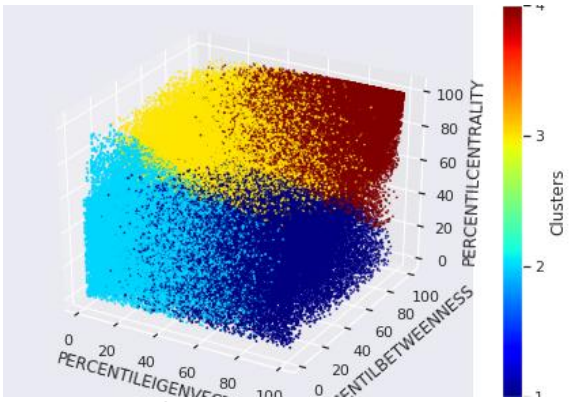
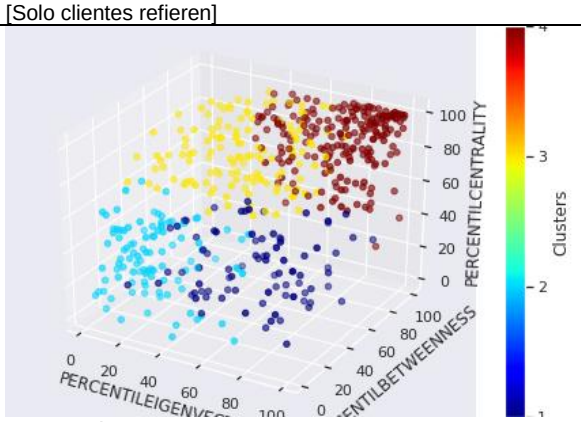

	

Ilustración 54: Scatterplot 3D variables E-B-C	*Ilustración 55: Scatterplot 3D variables E-B-C (Solo clientes refieren)*

Tal como se esperaba, se observa que los puntos pertenecientes al mejor conjunto suelen estar concentrados en los extremos más altos de cada métrica.

A partir de este análisis se puede concluir que, al estudiar percentiles, a diferencia de lo que ocurre con las métricas normalizadas, es posible clasificar a los clientes a nivel agregado.

Resulta importante mencionar que no es posible establecer cuales métricas son las que tienen un mayor efecto sobre la tasa de respuesta de un conjunto ni caracterizar a los clientes de cada uno, sin embargo, se observa que existe una correlación que permite describir el comportamiento a nivel agregado.

Resulta importante mencionar que se intentó utilizar variables demográficas y de vinculación para poder caracterizar a los clientes de cada conjunto, buscando la existencia de patrones o diferencias que permitan explicar su comportamiento. Sin embargo, debido a que estos datos corresponden a una campaña realizada durante el 2018, muchas de las variables utilizadas tenían una cantidad importante de missing values, no pudiendo obtener conclusiones verídicas de lo que ocurre en los conjuntos.

9.2 Desarrollo metodológico 2019

9.2.1 Construcción base datos relaciones

Para la construcción de la base de relaciones se realiza un procedimiento análogo al realizado sobre la base de datos anterior.

Se comienza seleccionando un subconjunto de la base de transferencias con datos desde el 01 de agosto de 2018 hasta el 30 de agosto de 2019. Se seleccionan las variables fecha y monto para describir la transacción, mientras que para describir al emisor y receptor se seleccionan rut y banco.

La tabla resultante contiene 40.490.584 **transferencias**, 2.884.318 personas **únicas** de las cuales 469.408 son clientes del banco y 2.414.910 no clientes.

Se caracteriza las **relaciones únicas** que existen en la base, tal como se aprecia a continuación:

Tipo relación		Numero relaciones únicas
Cliente	Cliente	283.896
Cliente	No Cliente	3.778.458
Total		4.062.354

Ilustración 56: Caracterización relaciones 2018/2019

A modo de análisis exploratorio, se estudia la distribución del monto y relaciones únicas de las transferencias.

El monto promedio de las transferencias en esta base es de $103.714, donde el 50% de las transferencias se realizan por montos de $29.000. En cuanto a las relaciones únicas se observa que en promedio dos personas se transfieren 4 veces durante el periodo analizado, a partir de los percentiles se observa que al menos la mitad de la base corresponde a relaciones que ocurren solo 1 vez.

El detalle de las distribuciones se puede encontrar en los anexos (Anexo sección 9.2.1), el descriptivo se observa a continuación:

	MONTO
count	34.346.408
mean	103.714
std	297.571
min	2
25%	10.000

	RELACIONES
count	8.122.389
mean	4
std	10
min	1
25%	1

50%	29.000
75%	100.000
max	180.000.000

Tabla 20: Descripción monto transferencias 2018/2019

50%	1
75%	3
max	2.629

Tabla 21: Descripción relaciones únicas transferencias 2018/2019

9.2.2 Construcción métricas relaciones

A partir de la base de transferencias se construye una base de **relaciones** con las mismas métricas que se habían calculado sobre la base de datos anterior.

- Delta
- Frequency
- Monto
- Recency
- RF

9.2.2.1 Análisis exploratorio métricas

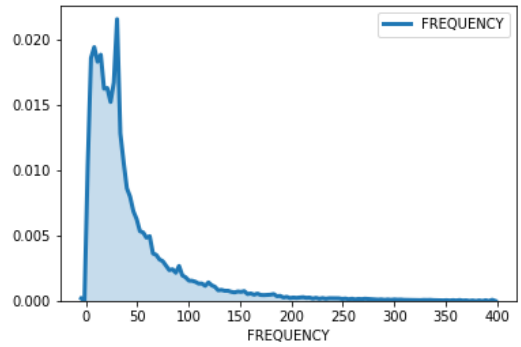
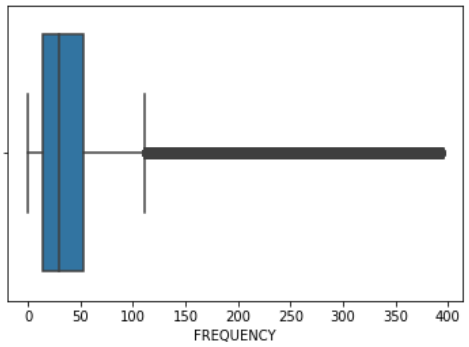
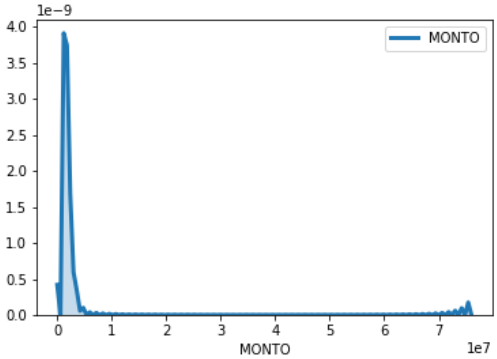
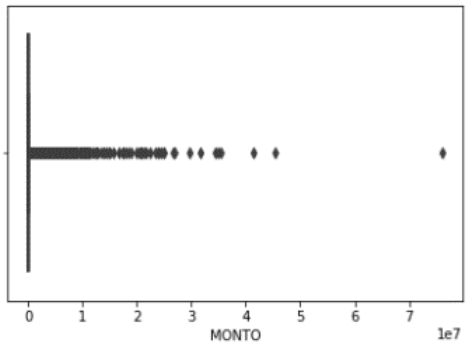
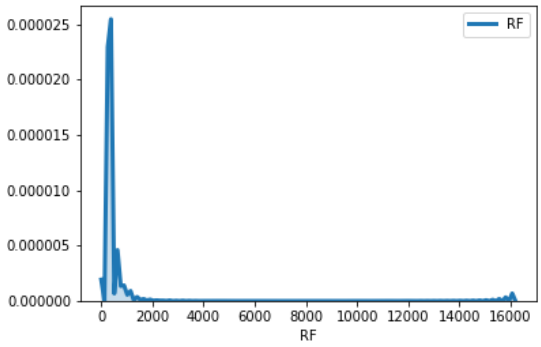
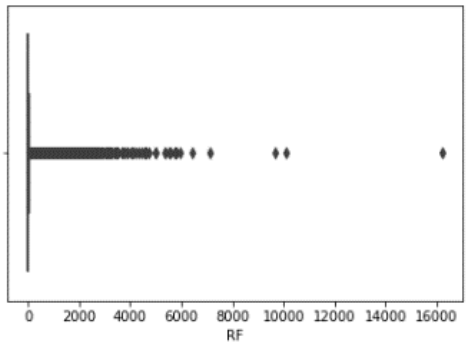
Resulta importante recordar que, dado que tanto Frequency como RF no pueden ser calculadas para relaciones que solo se dan 1 vez y considerando que una sola transferencia no es un buen indicador entre dos personas, se decide filtrar la base dejando solo aquellas relaciones en que se realizaron al menos dos transferencias.

La base resultante se compone de 4.062.354 relaciones. Sobre esta base se realiza un análisis exploratorio, que se muestra a continuación:

	DELTA	FREQUENCY	MONTO	RECENCY	RF
count	4.062.354	4.062.354	4.062.354	4.062.354	4.062.354
mean	159	44	90.127	104	12
std	126	50	216.844	102	52
min	1	0,02	500	0	0
25%	42	14	13.630	21	1
50%	129	29	31.500	63	2
75%	271	53	89.833	172	7
max	394	394	76.000.000	393	16.200

Tabla 22: Descriptivo tabla relaciones 2018/2019

Se estudia la distribución y Boxplot de las variables que tienen mayor presencia de outliers, el detalle del resto de las variables se encuentra en los anexos (Anexos 9.2.2.1)

1.1 Distribución variable Frequency:	1.2 Boxplot variable Frequency:
 *Ilustración 57: Distribución variable Frequency*	 *Ilustración 58: Boxplot variable Frequency*
2.1 Distribución variable Monto	2.2 Boxplot variable Monto
 *Ilustración 59: Distribución variable Monto*	 *Ilustración 60: Boxplot variable Monto*
3.1 Distribución variable RF:	3.2 Boxplot variable RF:
 *Ilustración 61: Distribución variable RF*	 *Ilustración 62: Boxplot variable RF*

En las ilustraciones se observa que, tal como mostraba en los descriptivos y como ocurrió con los análisis de la base anterior, las variables que están más concentradas en valores bajos y con una mayor presencia de outliers son Frequency, Monto y RF.

9.2.3 Limpieza de datos

Para la limpieza de datos nulos o outliers se procesa la base de relaciones analizando cada variable por separado y bajo el criterio de 3 desviaciones estándar. El detalle se muestra a continuación:

Variable	Media	Desviación Estándar	Límite Inferior	Límite Superior	Datos Filtrados
DELTA	159	126	-219	537	471.862
FREQUENCY	44	50	-106	194	100.676
MONTO	90.127	216.843	-560.405	740.658	56.009
RECENCY	103	102	-202	410	629.149
RF	12	52	-144	168	52.840

Tabla 23: Filtro outliers base relaciones 2018/2019

Eliminando los outliers, la base resultante se compone de 2.961.343 relaciones.

9.2.4 Modelamiento redes

EL modelamiento de redes tiene por objetivo construir la dimensión **nodo** de cada cliente. Se vuelve a recordar que para este análisis se utilizan algoritmos que iteran sobre la estructura de la red, por lo tanto, no se considera el "score" de cada relación, solo importa que una relación entre dos clientes exista, teniendo todas las relaciones igual peso o relevancia.

Se aplica un nuevo filtro iterativo sobre la base de relaciones para llegar a un tamaño de red que sea procesable dentro de los plazos que rigen este proyecto.

9.2.4.1 Filtro de red

La red inicial se compone de 2.359.883 nodos y 2.961.343 enlaces. El proceso iterativo se presenta a continuación:

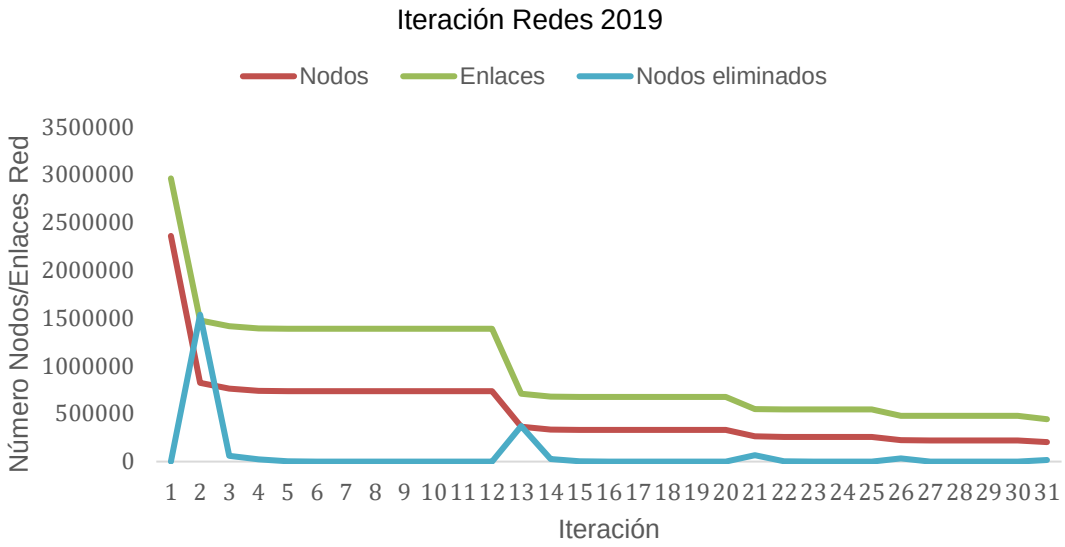
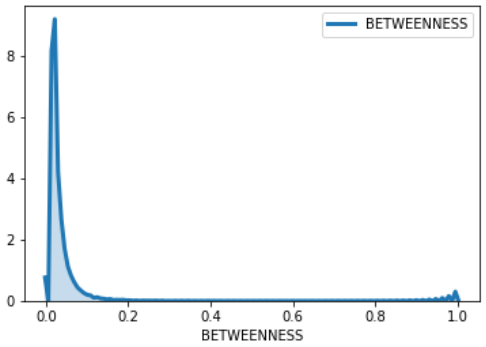
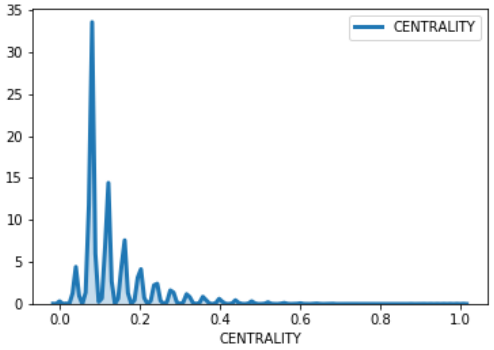


Ilustración 63: Iteración filtro de redes 2019

La red resultante se compone de 203.911 nodos y 443.210 enlaces. Del total de nodos en la red, 141.074 corresponde a clientes. El detalle de cada iteración se encuentra en los anexos (Anexo sección 9.2.4.1).

9.2.5 Métricas redes

Se analiza la distribución de las métricas calculadas sobre la red final. En primer lugar, se estudia la distribución de estas variables para chequear si tienen un comportamiento similar al estudiado en la red de 2018.

1. Distribución Betweenness	2. Distribución Centrality
 *Ilustración 64: Distribución Betweenness redes 2019*	 *Ilustración 65: Distribución Centrality redes 2019*
3. Distribución Eccentricity	4. Distribución Eigenvector

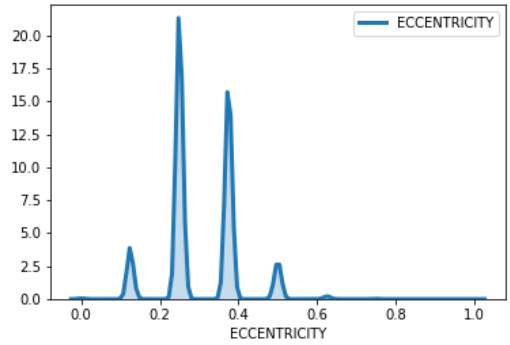


Ilustración 66: Distribución Eccentricity redes 2019

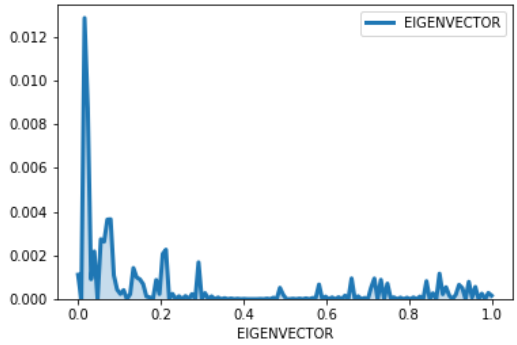


Ilustración 67: Distribución Eigenvector redes 2019

Se puede observar que tanto la variable Betweenness como Eigenvector están fuertemente concentradas en valores cercanos a 0, efecto que también se observaba en la red de 2018. Las variables Eccentricity y Centrality por ser números enteros, se concentran sobre un numero acotado de valores. El detalle de los gráficos Boxplot de estas variables, también estudiados en este análisis, se encuentra en los anexos (Anexos sección 9.2.5).

En base a lo ya estudiado, se considera que es adecuado para el modelo estudiar las métricas en forma de **percentiles** por lo que se realiza dicha transformación sobre cada una de ellas.

Finalmente, se construye la dimensión **nodo** promediando las métricas (en percentiles) para cada cliente.

9.2.6 Construcción dimensión enlace

A partir de la base de relaciones, se da paso a la construcción de la dimensión enlace de los clientes analizados.

9.2.6.1 Análisis e implementación modelos machine learning

Se aplican los modelos de Machine Learning, entrenados con la base de relaciones de 2017/2018, utilizando como datos de entrada aquellos de la base de relaciones de 2019.

Resulta importante recordar que los modelos se entrenaron con datos de relaciones proveniente de la base de transferencias para los años 2017 y 2018, utilizando como variable dependiente la campaña de referidos del 2018. En consecuencia, dado que se trata de modelos de clasificación, el resultado corresponde a la **probabilidad** que una relación entre un cliente y un no cliente se transforme en un referido.

Se recalca que más allá de lo que representa este resultado, se utiliza esta probabilidad como un cuantificador o “score” de cada relación.

Luego, a partir de este ponderador se crea la dimensión **enlace** de cada cliente. La que se construye promediando el valor de cada una de sus relaciones.

A partir de las métricas de desempeño analizadas en la sección anterior (9.1.5 Implementación modelos) se decide seleccionar los modelos regresión logística y Gradient Boosting por ser los más equilibrados en cuanto a las métricas de desempeño.

Dado lo similares que son estas métricas para ambos modelos se decide estudiar la distribución del vector de probabilidad que entregan como resultado. Esto ya que, como se mencionó anteriormente, los modelos se están utilizando para **clasificar** las relaciones y por lo tanto, para efectos de este análisis, la distribución que tendrá esta métrica para el conjunto de clientes estudiado resulta tan relevante como el score (probabilidad) de cada relación.

1. Logistic Regression	2. Logistic Regression Balanceado
Mean: 0.0002	Mean: 0.4274
Std: 0.00017	Std: 0.2043

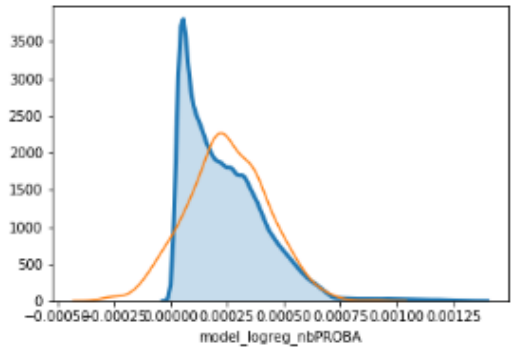
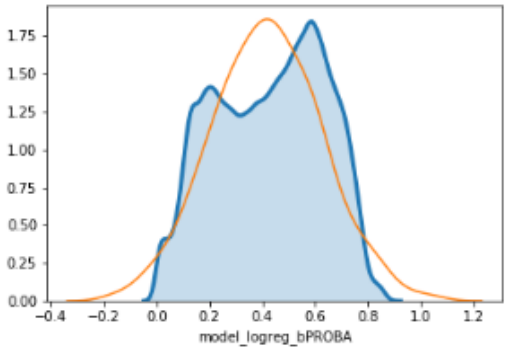

	

Ilustración 68: Comparación distribución normal con distribución Logistic Regression	*Ilustración 69: Comparación distribución normal con distribución Logistic Regression Balanced*
3. Gradient Boosting	4. Gradient Boosting Balanceado
Mean: 0.0053	Mean: 0.3779
Std: 0.0608	Std: 0.2749

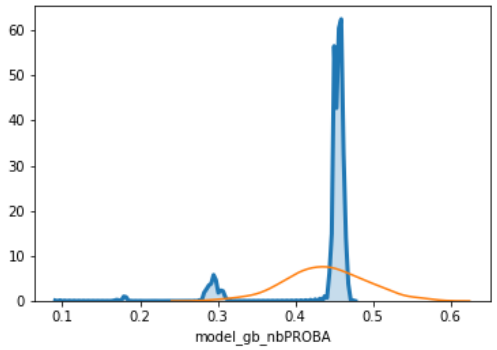
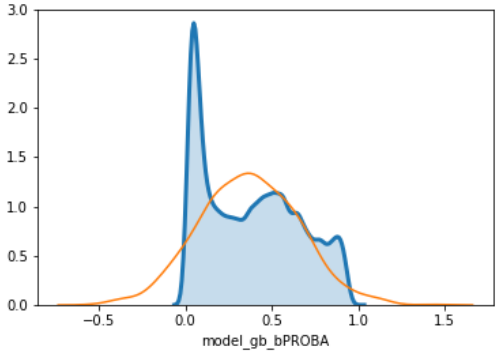

	

Ilustración 70: Comparación distribución normal con distribución Gradient Boosting	*Ilustración 71: Comparación distribución normal con distribución Gradient Boosting Balanced*

A partir de la comparación de las distribuciones de probabilidad entregada por cada modelo con una distribución normal, se distingue que la distribución de la regresión logística balanceada es la que mejor se ajusta a una normal, por lo tanto, se decide utilizar este modelo para el cálculo del “score” de cada relación. Esta característica se considera necesaria ya que se busca discernir entre la “calidad” de las relaciones más que concluir directamente sobre el valor que el modelo le asigna a cada relación.

Finalmente, se construye la dimensión **enlace** promediando los “score” de las relaciones de cada **cliente**. Luego este cálculo se lleva a percentiles para estar en línea con el orden de magnitud de la dimensión nodo antes calculada.

9.3 Construcción relevancia cliente

Una vez calculadas ambas dimensiones que indican la relevancia de un cliente, **nodo** y **enlace**, se realiza el cruce entre ambas tablas dejando solo a los clientes (recordando que en la tabla de relaciones también contiene no clientes) para los siguientes análisis.

9.4 Piloto campaña referidos

Se realiza un piloto correspondiente a una campaña de referidos. Esto se hace con el objetivo de contrastar las métricas calculadas para determinar si estas influyen en que un cliente refiera o no.

La campaña consiste en entregar un incentivo (o gancho) en puntos por cada apertura que un cliente genera. Los clientes deben ingresar a un formulario online indicando quienes son las personas que están refiriendo para que obtengan su tarjeta de crédito con el banco.

9.4.1 Diseño experimental

9.4.1.1 Clientes

Los clientes seleccionados para recibir la campaña provienen de dos orígenes. El primer conjunto corresponde a los clientes analizados en la red, mientras que el segundo conjunto corresponde a un grupo de control de 100.000 personas seleccionadas al azar dentro de la cartera completa de clientes del banco.

Resulta importante recordar que, como se mencionó anteriormente, el grupo de control tiene por objetivo contrastar el valor de haber identificado a los clientes relevantes. En efecto, ambos grupos (tratamiento y control) reciben exactamente la misma campaña.

Por otro lado, se debe tener en consideración que, dado que los clientes que se analizan de la red provienen de la tabla de transferencias, en su mayoría son clientes que poseen cuenta corriente por sobre tarjeta de crédito. Esta proporción no se condice con la proporción de clientes que provienen del conjunto aleatorio los que en su mayoría son

clientes solo tarjeta de crédito. El detalle de estas proporciones se encuentra a continuación:

	Clientes cuenta corriente	Clientes tarjeta crédito	Clientes ambos productos	Total
Grupo tratamiento (Redes)	48.601 (34%)[19]	20.680 (15%)	71.793 (51%)	141.074 (100%)
Grupo control (Aleatorios)	15.784 (16%)[20]	66.364 (66%)	17.852 (18%)	100.000 (100%)

Tabla 24: Caracterización clientes analizados piloto

9.4.1.2 Gancho

El gancho ofrecido en la campaña es de 3.000 puntos por apertura llegando a un máximo de 30.000 puntos por cliente. El valor o costo de cada punto se puede aproximar[21] a $2.5 por punto, en consecuencia, el gancho tiene un valor de $7.500 por apertura con tope de asignación de $75.000 por cliente.

9.4.1.3 Contactos

Debido a que para efectos del proyecto es fundamental que los clientes estén conscientes de la existencia de la campaña, esta se comunica de forma multicanal siendo email el principal medio de comunicación, pero incluyendo también banner en la aplicación móvil y banner en la página web del banco.

9.4.1.4 Formulario

El mail enviado a los clientes con la campaña contiene una URL al formulario donde se ingresan las personas referidas. El que se presenta a continuación:

[19] Porcentaje calculado respecto al total del grupo tratamiento
[20] Porcentaje calculado respecto al total del grupo de control
[21] Valor corresponde a una aproximación debido a que el valor del punto depende de la categoría a la que pertenece el cliente.

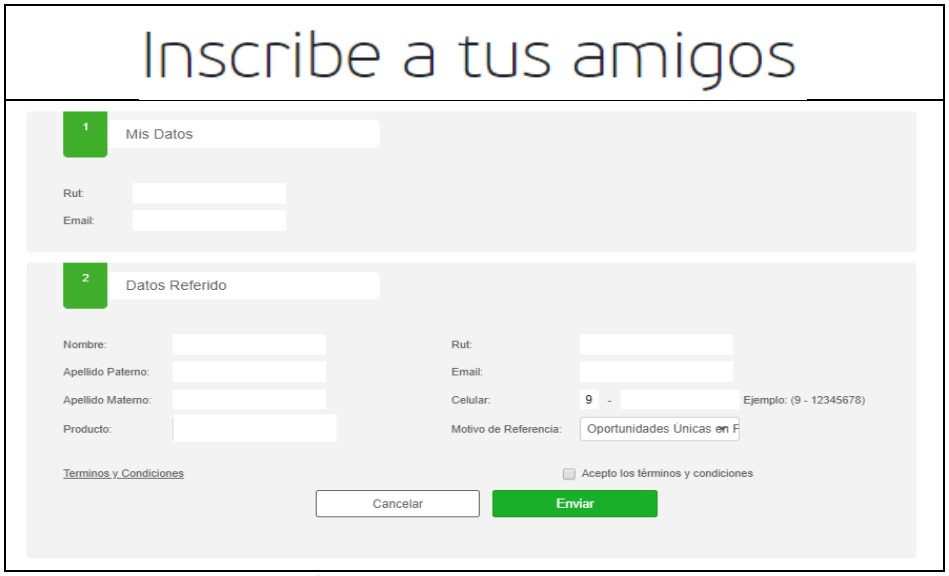


Ilustración 72: Formulario referidos campaña 2019

9.4.1.5 Graficas campaña

Se presentan las gráficas utilizadas para la campaña:

- **Graficas email**

Ilustración 73: Graficas email campaña 2019

- **Grafica banner aplicación móvil**

Ilustración 74: Grafica email campaña 2019

9.4.1.6 Funnel campaña

Existen distintos motivos que limitan la contactabilidad de las campañas del banco, por ejemplo, clientes que hayan presentado solicitudes para no recibir promociones en sernac, clientes que hayan marcado al emisor como spam o clientes que pertenecen al grupo de "no contactables" del banco, estos últimos no reciben ninguna campaña durante un año con el objetivo de evaluar estrategias comerciales del banco.

Se estudia el Funnel de la campaña para analizar, del conjunto de clientes seleccionados, cuantos efectivamente se consideran como participantes de la campaña, el detalle se presenta a continuación:

Clientes	Seleccionado para piloto	Recibe mail campaña	Abre mail campaña	Refiere
Cliente Red	141.074 (100%)	131.365 (93%)	39.588 (30%)	332 (0,8%)
Cliente Aleatorio	100.000 (100%)	91.114 (91%)	17.962 (20%)	54 (0,3%)
Cliente ambos conjuntos	4.250 (100%)	4.073 (95%)	1.213 (30%)	8 (0,7%)

Tabla 25: Funnel campaña referidos 2019

Resulta importante fijar el criterio para definir el universo de clientes a utilizar pensando en que esto puede tener implicancias sobre la tasa de respuesta utilizada para los análisis posteriores.

Se estudia cual es la condición que abarca la mayor cantidad de clientes que hayan referido gente en la campaña.

Referidos	Recibe mail campaña	Abre mail campaña	Total
Cliente Red	330 (99%)	260 (78%)	332 (100%)
Cliente Aleatorio	54 (100%)	47 (87%)	54 (100%)
Cliente ambos conjuntos	8 (100%)	6 (75%)	8 (100%)

Tabla 26: Clientes que contactados por categoría

A partir de la tabla se logra apreciar que al menos el 99% de los clientes que participa refiriendo gente recibe el email con la campaña. Por lo tanto, se define este como parámetro para fijar el universo de clientes en los análisis posteriores.

Una vez filtrado el universo de clientes, la distribución separando por origen y por tipo de producto que poseen con el banco queda:

Numero Clientes	Cliente tarjeta crédito	Cliente cuenta corriente	Cliente Mixto	Total
Cliente Red	18.637	42.550	66.602	127.789
Cliente Aleatorio	50.876	9.461	13.647	73.984
Cliente ambos conjuntos	552	1.363	2.048	3.963
Total	70.065	53.374	82.297	205.736

Tabla 27: Clientes que se analizan por categoría

La distribución de clientes que participan refiriendo en la campaña queda:

Clientes Refieren	Cliente tarjeta crédito	Cliente cuenta corriente	Cliente Mixto	Total
Cliente Red	27	61	242	330
Cliente Aleatorio	35	4	15	54
Cliente ambos conjuntos	4	0	4	8
Total	66	65	261	392

Tabla 28: Clientes que refieren por categoría

Finalmente, el caculo de la tasa de respuesta de los conjuntos es directo sobre las dos tablas anteriores, como se detalla a continuación:

Tasa Respuesta	Cliente tarjeta crédito	Cliente cuenta corriente	Cliente Mixto	Total
Cliente Red	0,14%	0,14%	0,36%	0,26%
Cliente Aleatorio	0,07%	0,04%	0,11%	0,07%
Cliente ambos conjuntos	0,72%	0,00%	0,20%	0,20%
Total	0,09%	0,12%	0,32%	0,19%

Tabla 29: Tasa de respuesta por categoría

9.5 Análisis estadístico

Este análisis busca establecer si existe evidencia estadística para afirmar que las variables que indican la relevancia de un cliente están describiendo la tasa de respuesta de la campaña y este efecto no es derivado del azar.

Para lo anterior, se construye una variable dependiente análogo a lo realizado en la campaña del 2018, pero ahora estudiado a nivel de cliente. La variable se define a continuación:

$$Refiere_i = \begin{cases} 1\ Si\ cliente\ i\ refiere\ en\ la\ campaña\ 2019 \\ 0\ No\ refiere \end{cases}$$

Ecuación 8: Definición de variable dependiente Refiere 2019

El efecto se estudia por medio de las dos dimensiones con que busca explicar la relevancia de cada cliente, **nodo** y **enlace** en conjunto con variables demográficas y de vinculación.

Dentro de las variables demográficas se tiene:

- Edad: se estudia de forma estandarizada, es decir, valores entre 0 y 1.
- Género: se define como 1 para hombres, 0 mujeres.
- Número de hijos: se estudia de forma normalizada.
- Casado: se define como 1 personas casadas, 0 solteros.

Dentro de las variables de vinculación se tiene:

- Enganchado Promoción: variable binaria que indica si cliente suele reaccionar a promociones con la tarjeta de crédito.
- Enganchado Puntos: variable binaria que indica si el cliente canjea puntos de forma recurrente.
- Enganchado Financiamiento: variable binaria que indica si cliente utiliza la tarjeta como financiamiento, es decir, pago en cuotas, avance de tarjeta, etc.
- Principalidad: Variable categórica que mide el nivel de vinculación de un cliente con el banco en base al uso de productos y cruce de productos. Puede tomar 4

valores: Muy alta, alta, media y baja. Para los análisis posteriores, se utiliza la categoría baja como caso base de comparación.

9.5.1 Análisis correlación

Resulta importante mencionar que los análisis posteriores se realizan sobre la relevancia de los clientes, por lo tanto, se debe excluir a los clientes del grupo aleatorio ya que no se cuenta con métricas que describan su relevancia.

A modo de análisis exploratorio, se observa la correlación de las variables independientes con la variable Refiere, el detalle se observa a continuación:

	CORRELACIÓN REFIERE
Principalidad Muy alta	1,94%
Enganchado Promoción	1,01%
Enganchado Puntos	0,88%
Nodos	0,88%
Enlaces	0,85%
Enganchado Financiamiento	0,44%
Sexo	0,27%
Principalidad Alta	-0,61%
Principalidad Media	-1,33%
Número Hijos	-1,83%
Casado	-1,90%
Edad	-2,98%

Tabla 30: Correlación variable dependiente campaña referidos 2019

De la tabla se observa que las variables de vinculación principalidad muy alta y enganchado promoción son las que tienen mayor correlación positiva con la variable dependiente. De mayor magnitud, pero en sentido contrario, se encuentran las variables demográficas casado y edad.

9.5.2 Regresiones logísticas

Se testean las hipótesis propuestas utilizando regresiones logísticas sobre distintos conjuntos de variables para evaluar el poder explicativo de cada una, su significancia estadística y la bondad de ajuste de los modelos planteados.

Resulta importante volver a recordar que, dado que se trata de regresiones logísticas, la interpretación de las variables se hace sobre su Odds ratio y no sobre el coeficiente.

Las regresiones se ejecutan sobre el 80% de los datos para ser evaluadas en el 20% restante.

- Regresion logistica (Todas las variables)

Variable	Coefficient	Odds	p-value
Intercept	-4,969 (***)	0,007	0,000
NODO	0,215	1,239	0,471
ENLACE	0,245	1,277	0,317
EDAD	-4,178 (***)	0,015	0,000
N_HIJOS	-0,411	0,663	0,739
CASADO	-0,491 (*)	0,612	0,024
GENERO	-0,043	0,958	0,755
ENGANCHADO PUNTOS	0,209	1,232	0,158
ENGANCHADO FINANCIAMIENTO	0,178	1,195	0,349
ENGANCHADO PROMOCION	0,287 (*)	1,333	0,041
MUYALTA	0,488	1,629	0,227
ALTA	-0,173	0,841	0,709
MEDIA	-0,488	0,614	0,333

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Tabla 31: Regresión logística campaña 2019

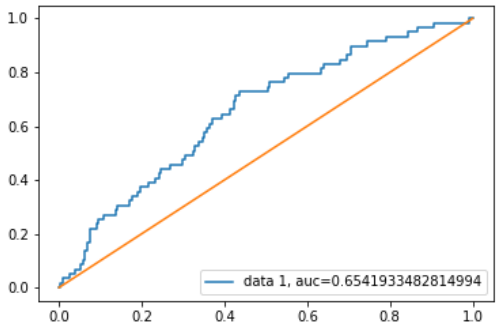


Ilustración 75: Curva ROC Regresión logística campaña 2019

De la regresión se observa que las variables Nodo y enlaces (dimensiones de **relevancia** de cada cliente) muestran una relación positiva con la variable dependiente, es decir, en promedio, un cliente con mayor calidad de enlace y nodo tiene una mayor tasa de respuesta a la campaña. Si bien el efecto está en línea con lo esperado, no es posible afirmar dicha relación debido a que ninguna de las dos variables es estadísticamente significativa.

Respecto a las variables demográficas se obtiene que edad y casado son significativas. Ambas con Odds ratio menor que 1, por lo tanto, en promedio, un cliente de mayor edad tendría un desempeño inferior que aquellos más jóvenes. En el mismo sentido se concluye respecto a los clientes casados versus aquellos solteros.

En cuanto a las variables de “vinculación” se observa que las magnitudes van en el sentido esperado. Para las variables “Enganchado” se tiene Odds ratio mayores que 1 lo que denota una relación positiva con la variable dependiente. La variable enganchado promoción es la única variable significativa lo que se encuentra en línea con el gancho en puntos de la campaña.

En cuanto a las variables de principalidad, se tiene que ninguna variable es significativa. Los Odds ratio son positivos y van descendiendo en magnitud a medida que se baja en categoría. Recordando que el caso base de la variable corresponde a la categoría principalidad baja, se esperaría que todos los coeficientes fueran mayores que 1, sin embargo, esto solo ocurre para la principalidad muy alta.

Se realiza una segunda regresión logística, esta vez sin las variables “Enganchado” debido a que por la cantidad de missing values, utilizarlas implica perder un 30% de los

clientes que refieren lo que podría estar afectando la magnitud o significancia de los variables analizadas.

- Regresion logistica (Sin variables Eganchado)

Variable	Coefficient	Odds	p-value
Intercept	-5,310 (***)	0,005	0,000
NODOS	0,240	1,272	0,368
ENLACES	0,174	1,190	0,431
EDAD	-3,606 (***)	0,027	0,000
N_HIJOS	-0,357	0,700	0,735
CASADO	-0,298	0,742	0,105
GENERO	-0,251 (*)	0,778	0,047
MUYALTA	0,867 (*)	2,380	0,011
ALTA	0,347	1,415	0,346
MEDIA	-0,523	0,593	0,261

Tabla 32: Regresión logística campaña 2019 (Sin variables Enganchado)

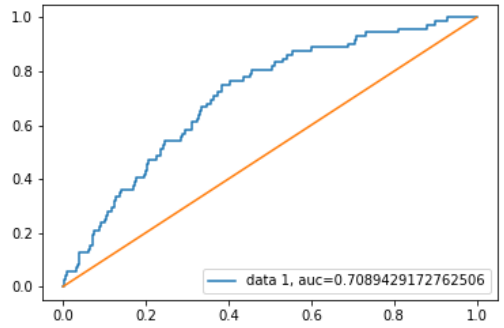


Ilustración 76: Curva ROC Regresión Logística campaña 2019 (sin variables Enganchado)

Se observa que las variables de relevancia de los clientes varían en magnitud de Odds ratio, pasando la variable Nodo a tener el mayor efecto entre estas dos. Variación similar, pero en sentido inverso ocurre con el p-valor de las variables, aun así, ambas continúan lejos de ser significativas.

Respecto a las variables demográficas, se tiene que genero pasa a ser significativa con un Odds ratio muy cercano a 0, recordando que el caso base corresponde a mujeres, se tiene que, en promedio, las mujeres tienen un mayor impacto sobre la variable dependiente.

Respecto a las variables de vinculación (ahora solo principalidad), se tiene que la principalidad muy alta pasa a ser significativa y aumenta en magnitud de Odds ratio, lo que se podría explicar por estar absorbiendo parte del efecto que era explicado por las variables enganchado.

Se concluye que, a nivel de clientes, con las variables analizadas no es posible generar con un modelo con la bondad de ajuste suficiente para determinar si existe un efecto estadísticamente significativo de las variables calculadas para explicar la relevancia de un cliente bajo el contexto de generar aperturas en una campaña de referidos.

A partir de lo anterior se da paso a un análisis no supervisado para estudiar si es posible determinar un efecto a nivel agrupado.

9.6 Análisis no supervisado

Se realiza un análisis no supervisado en base a las dos métricas que describen la relevancia de cada cliente, **nodo** y **enlace**.

Se comienza utilizando dos métodos para determinar el número de clúster a utilizar, Elbow y Calinski Harabasz.

- Elbow Method

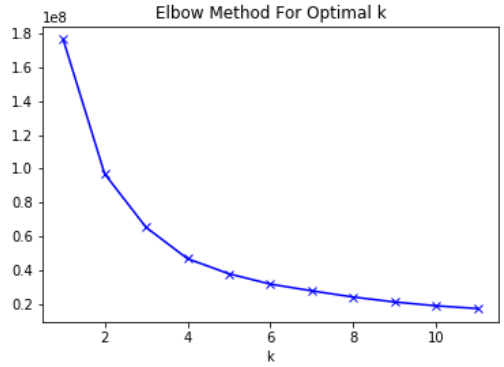


Ilustración 77: Elbow curve datos campaña 2019

- Calinski Harabasz Method

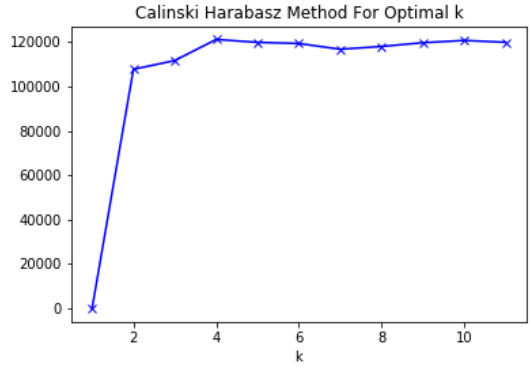


Ilustración 78: Calinski Harabasz curve datos campaña 2019

Se decide fijar el número de clúster en 4, esto debido a que ambas curvas coinciden en este valor. Por un lado, el método Calinski Harabasz alcanza un máximo en 4, mismo punto donde la curva Elbow muestra una variación en su pendiente.

Se realiza una agrupación de los valores para estimar, la tasa de respuesta promedio de cada conjunto, el número de clientes en cada clúster y el valor de cada métrica para los centroides.

clúster	ENLACES	NODOS	REFIERE	Personas
4	72,3	71,8	0,35%	33.367
3	28,5	71,8	0,24%	31.251
2	78,8	32,8	0,23%	34.541
1	24,6	33,5	0,21%	32.593

Ilustración 79: Agrupación clúster campaña 2019

Se observa que el centroide del conjunto con mejor tasa de respuesta (clúster 4) posee el máximo valor en ambas métricas de relevancia. Por otro lado, se observa que la distribución de clientes es homogénea en los 4 conjuntos lo que aporta a la confianza de esta clasificación.

Se calcula un test de diferencia en proporciones para determinar si la diferencia en las tasas de respuesta de los grupos es estadísticamente significativa.

clúster	4	3	2	1
4	0	2.52 (95%)	2.82 (95%)	3.47 (99%)
3		0	0.22	0.92
2			0	0.72
1				0

Valores en paréntesis representan significancia estadística
Tabla 33: Test diferencia en proporciones clúster campaña 2019

Se observa que la tasa de respuesta del clúster 4 es estadísticamente superior a los otros 3 conjuntos. Sin embargo, la diferencia entre estos 3 conjuntos no es significativa, por lo tanto, no es posible determinar si la diferencia observada en las tasas de respuesta es producto de la clasificación o se debe a un hecho aleatorio.

Se realiza un análisis de dispersión para ver si es posible distinguir las agrupaciones de los conjuntos y su relación con la tasa de respuesta. El grafico se muestra a continuación:

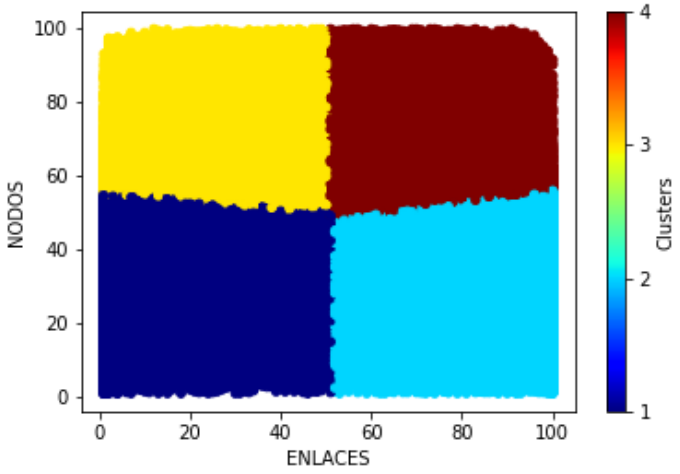


Ilustración 80: Análisis de dispersión variables de relevancia clientes campaña 2019

A partir de la gráfica se observa que el conjunto con mejor tasa de respuesta (clúster 4) se concentra en valores altos tanto para la dimensión nodo como enlace. El segundo mejor conjunto (clúster 3) se agrupa con valores altos para la dimensión nodo, pero bajos para la dimensión enlace.

A partir de este gráfico se observa que la dimensión nodo está siendo de mayor relevancia para explicar la tasa de respuesta de la campaña, ya que los dos mejores conjuntos muestran valores altos en esta variable. Esto se encuentra en línea con la regresión logística estudiada sin las variables enganchado donde el Odds ratio de la variable nodo era superior al de la variable enlace.

Se busca complementar el grafico anterior con un mapa de calor a nivel de deciles. Con esto se busca determinar si los conjuntos están siendo homogéneos dentro de sí o los

efectos observados se deben a un segmento acotado de clientes que marcan una tendencia para todo su conjunto.

- **Mapa Calor tasa respuesta por deciles**

Nodos										
10	0,00%	0,20%	0,30%	0,40%	0,10%	0,10%	0,30%	0,10%	1,10%	0,70%
9	0,20%	0,40%	0,20%	0,20%	0,30%	0,30%	0,30%	0,70%	0,20%	0,80%
8	0,50%	0,10%	0,10%	0,50%	0,40%	0,50%	0,20%	0,10%	0,30%	0,70%
7	0,10%	0,10%	0,30%	0,50%	0,30%	0,20%	0,40%	0,30%	0,30%	0,60%
6	0,10%	0,20%	0,40%	0,20%	0,10%	0,30%	0,20%	0,40%	0,30%	0,20%
5	0,10%	0,20%	0,20%	0,10%	0,20%	0,20%	0,30%	0,30%	0,40%	0,30%
4	0,30%	0,30%	0,20%	0,20%	0,20%	0,40%	0,20%	0,10%	0,20%	0,10%
3	0,20%	0,10%	0,20%	0,30%	0,20%	0,10%	0,20%	0,10%	0,10%	0,10%
2	0,10%	0,40%	0,30%	0,10%	0,00%	0,40%	0,10%	0,10%	0,50%	0,40%
1	0,50%	0,00%	0,00%	0,00%	0,40%	0,00%	0,00%	0,40%	0,60%	0,30%
Enlaces	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Tabla 34: Mapa calor tasa respuesta

- **Mapa calor clientes por deciles**

Nodos										
10	185	470	627	732	840	902	889	730	636	286
9	495	976	1.275	1.403	1.529	1.472	1.517	1.410	1.144	663
8	842	1.464	1.558	1.745	1.805	1.856	1.742	1.753	1.593	1.132
7	1.179	1.688	1.772	1.748	1.817	1.842	1.941	1.941	1.804	1.611
6	1.563	1.844	1.889	1.890	1.841	1.928	1.853	1.882	1.927	1.993
5	1.668	1.948	1.891	1.809	1.721	1.904	1.881	1.904	2.064	2.299
4	1.563	1.690	1.658	1.641	1.553	1.563	1.588	1.796	1.872	2.332
3	1.289	1.293	1.268	1.187	1.256	1.230	1.281	1.383	1.434	2.156
2	965	807	757	728	781	797	788	804	965	1.601
1	412	250	242	212	225	201	223	267	342	639
Enlaces	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Tabla 35: Mapa calor número clientes

Respecto al primer mapa de calor, se observa que las mayores tasas de respuesta se encuentran concentradas en los 2 deciles más altos tanto para nodos como para enlaces. Mientras que las tasas de respuesta menores se concentran en el decil más bajo en la variable nodos. Sin embargo, en aspectos generales no es posible distinguir una tendencia clara que permita relacionar la tasa de respuesta con las métricas estudiadas.

Respecto al segundo mapa de calor, se tiene que para la concentración de clientes si se observa una tendencia en línea con lo esperado. En rasgos generales, se ve una mayor concentración de personas en los deciles centrales para ambas métricas.

10 RESULTADOS

Los resultados son analizados siguiendo el enfoque definido anteriormente, es decir, la tasa de respuesta a las campañas de referidos se analiza mediante el número de clientes que participa en la campaña de referidos.

10.1.1 Efecto incremental

El efecto incremental busca medir el impacto de haber recibido el tratamiento, esto se lleva a cabo comparando al grupo tratado con el grupo de control. Para efectos de este proyecto, a diferencia de lo que normalmente se interpreta como efecto incremental, el impacto medido no está relacionado al hecho de recibir o no el tratamiento, sino al hecho de haber sido identificado como cliente relevante.

Como se ha mencionó anteriormente, el grupo de control (segmento aleatorio) recibió exactamente el mismo tratamiento que el grupo tratado (segmento analizado mediante redes). En la tabla se exhiben las tasas de respuesta los distintos clústers identificados y del grupo de control.

clúster	REFIERE [Tasa respuesta]	Personas
4	0,35%	33.367
3	0,24%	31.251
2	0,23%	34.541
1	0,21%	32.593
Control	0,09%	77.947

Tabla 36: Tasa de respuesta segmentos campaña

Se realiza un test de diferencia en proporciones para determinar si las diferencias de la tasa de respuesta entre cada clúster son estadísticamente significativas.

clúster	4	3	2	1	Control
4	0	2.52 (95%)	2.82 (95%)	3.47 (99%)	16.38 (99%)
3		0	0.22	0.92	13.56 (99%)
2				0.72	13.33 (99%)
1				0	12.54 (99%)
Control					0

Tabla 37: Test de diferencia en proporciones clúster y grupo de control

De la tabla se observa que, para el grupo de control, todas las diferencias son estadísticamente significativas, sin embargo, al mirar las diferencias entre los distintos clústers se observa que solo son estadísticamente significativas las diferencias del clúster 4 con los otros conjuntos. Dado que la tasa de respuesta de este conjunto es superior a todo el resto, en particular al grupo de control, es posible calcular el efecto incremental del modelo como la diferencia entre las aperturas potenciales y reales. Se entiende por aperturas potenciales a la obtenida de extrapolar la tasa de respuesta del mejor clúster sobre la cantidad de clientes en el grupo de control.

Del análisis anterior se obtiene que, del total de clientes del grupo de control, 67 clientes participan en la campaña refiriendo a otra persona. Al extrapolar la tasa de respuesta del conjunto 4 sobre el grupo de control se logra que un total de 259 clientes participen en la campaña de referidos. Se obtiene entonces un efecto incremental de 192 clientes.

10.1.2 Ajuste proporcional

A partir de los análisis realizados anteriormente se observa que la tasa de respuesta del grupo de control es muy inferior a la tasa de respuesta del clúster con peor desempeño. Esta diferencia podría ser atribuible a la forma en que se construyen los conjunto por lo que resulta importante limpiar este efecto para poder mejorar la calidad de los resultados.

Como una primera aproximación a este delta se debe observar que existe una notoria diferencia en cómo se distribuyen los clientes según los productos contratados (tarjeta de crédito, cuenta corriente o ambos) entre el grupo de control y los clientes analizados en las redes. Se observa que el grupo de control, por tratarse de un segmento aleatorio de clientes, tiene proporciones representativas de la distribución que siguen los clientes del holding, siendo la mayoría clientes tarjeta de crédito. Por otro lado, los clientes que se analizan en las redes, por construcción se componen principalmente por clientes con cuenta corriente, debido a que las redes se obtienen del análisis de transferencias bancarias. Esta discrepancia se puede apreciar de forma gráfica en la siguiente tabla.

Numero Clientes	Cliente tarjeta crédito	Cliente cuenta corriente	Cliente Mixto	Total
Cliente Red	15%	33%	52%	127.789
Cliente Aleatorio	69%	13%	18%	73.984
Cliente ambos conjuntos	14%	34%	52%	3.963

Tabla 38: Distribución productos contratados clientes segmentados por origen

Esta diferencia podría estar alterando los resultados debido a la posible correlación entre la vinculación de un cliente y los productos contratados con él banco. A partir de lo anterior, se realiza un ajuste a las proporciones de los clientes en red seleccionando un subconjunto aleatorio de clientes e imponiendo una distribución resultante similar a la de los clientes del grupo de control. Las nuevas distribuciones se muestran a continuación.

Numero Clientes	Cliente tarjeta crédito	Cliente cuenta corriente	Cliente Mixto	Total
Cliente Red	19.312	3.591	5.180	28.083
Cliente Aleatorio	50.876	9.461	13.647	73.984
Cliente ambos conjuntos	572	106	153	832

Tabla 39: Distribución ajustada de productos contratados segmentados por origen

Esta transformación también repercute en la tasa de respuesta de los distintos clústers estudiados en el análisis de redes. La respuesta de cada conjunto y el número de clientes se muestra a continuación.

clúster	REFIERE [Tasa respuesta]	Personas	% Clientes cuenta corriente	% Clientes tarjeta crédito
4	0,33%	6570	42%	83%
3	0,18%	6119	36%	85%
2	0,16%	7588	27%	89%
1	0,11%	7944	25%	90%
Red	0,19%	28221	32%	87%
Control	0,12%	38981	38%	88%

En el análisis se incluye la fila Red, la que representa el promedio de las métricas de clientes analizados en los clústers

Tabla 40: Tasa de respuesta clúster ajustados

Se observa que con la nueva clasificación se conserva un numero equitativo de clientes en los distintos conjuntos. A la vez, que se tienen valores más homogéneos en las tasas de respuesta. Aun así, es posible distinguir que el clúster identificado como más relevante (por las métricas de nodo y enlace) corresponde al conjunto con mejor tasa de respuesta. Se estudia además la proporción de clientes banco y tarjeta de crédito para cada conjunto, siendo el orden de magnitud de estas proporciones similar entre los clústers y sin mayores discrepancias respecto al grupo de control.

Se realiza un test de diferencia en proporciones para discernir si la diferencia en la tasa de respuesta de los distintos clústers es estadísticamente significativa.

clúster	4	3	2	1	Control
4	0	0.08	0.10	0.29	0.02
3		0	0.03	0.21	0.08
2			0	0.17	0.12
1				0	0.41
Control					0

Tabla 41: Test de diferencia en proporciones con segmentos ajustados

Se observa que las diferencias entre los distintos conjuntos no son estadísticamente significativas, por lo tanto, no es posible afirmar que existe un mejor desempeño del clúster 4, compuesto por los clientes más relevantes, sobre el resto del universo estudiado.

10.2 Caracterización clientes

A partir de las variables de vinculación y demográficas analizadas anteriormente, se realiza una caracterización de los clientes que participan en la campaña para chequear si las variables utilizadas para estos análisis permiten generar una caracterización de los clientes identificados como “relevantes”.

Se realiza una primera caracterización sobre el universo de clientes que participa en la campaña, es decir, aquellos que reciben la comunicación por correo. Para los 205.736 clientes se tiene que la edad promedio es de 31 años, respecto a la variable género se observa que un 36% de la base son mujeres y el restante 64% hombres. De acuerdo con las variables de vinculación se tiene que un 51% se caracteriza como cliente que reacciona a ofertas del programa de fidelización, un 38% lo hace a promociones y un 85% ha contratado productos de financiamiento con el banco. De acuerdo con el indicador de principalidad, se tiene que un 26% posee principalidad alta, un 23% principalidad media y un 32% principalidad baja.

Se observan importantes diferencias tanto para las variables demográficas como de vinculación al momento de caracterizar a los clientes identificados como relevantes (pertenecientes al clúster con mejor tasa de respuesta). En primer lugar, para los 6.570 clientes se observa una edad promedio superior a la vista anteriormente llegando a 33 años, la distribución de genero se invierte pasando a ser un 66% mujeres y un 34% hombres. Respecto a las variables de vinculación se tiene que 49% reacciona a ofertas de fidelización, un 25% lo hace ante promociones mientras que un 7% a contratado algún producto de financiamiento con el banco. El indicador de principalidad muestra una concentración mayor para los valores altos, siendo un 36% clientes de principalidad alta, un 26% de principalidad media y solo un 22% posee una principalidad baja.

11 JUSTIFICACIÓN ECONÓMICA

Dado que el banco pertenece a un holding con distintas unidades de negocio y que la mayoría de los clientes participan en más de uno de estos, el enfoque planteado en este proyecto se podría implementar también en los otros comercios. Esto podría generar beneficios económicos derivados de una reducción de costos, por ejemplo, en campañas de fidelización, así como un aumento en los ingresos a causa de mejores campañas para los distintos productos ofrecidos por cada uno de los negocios.

No obstante, la estimación económica del proyecto se realiza solo para el banco, considerando la implementación de campañas de apertura, en particular campañas de referidos, aplicadas sobre los 4 principales productos del banco, estos son, apertura de tarjeta de crédito, super avance, avance y crédito de consumo.

El análisis ser realiza considerando los siguientes supuestos:

Producto	Rentabilidad promedio anual	Porcentaje	Universo clientes
Tarjeta crédito	$50.000	85%	2.975.000
Super avance	$30.000	12%	420.000
Avance	$15.000	12%	420.000
Crédito consumo	$20.000	5%	175.000

**Los datos de rentabilidad corresponden a un dato ficticio por tratarse de información sensible.*

Tabla 42: Supuestos estimación económica

A partir de lo anterior, se propone la implementación de una campaña de referidos al año para cada producto, por un periodo total de 5 años. Para cada campaña se contrastan las aperturas generadas según dos escenarios, un escenario base que intenta emular como se realizaría una campaña en el contexto actual y un escenario propuesto en el que se busca reflejar una campaña realizada utilizando la implementación de redes de clientes. Para el escenario base se considera la implementación de la campaña sobre el total de clientes del banco y la tasa de respuesta promedio obtenida del piloto de este proyecto. El segundo escenario se construye bajo la realización de la campaña solo sobre los clientes identificados como relevantes y considerando la tasa de respuesta obtenida por este conjunto en el piloto. El detalle de los cálculos se encuentra en los anexos.

A partir de esta estimación se obtiene un delta de aperturas entre el escenario utilizando redes versus el escenario base. Este delta se multiplica por la rentabilidad promedio de un cliente obteniéndose flujos de dinero anuales. Debido a que se trata de flujos futuros, es necesario traerlos a valor presente, para esto se utiliza una tasa de descuento de un 12%.

Producto	Año 1	Año 2	Año 3	Año 4	Año 5
Tarjeta crédito	$2.648.185	$2.598.722	$2.698.719	$2.548.729	$2.848.714
Super avance	$198.552	$168.596	$228.594	$138.600	$318.591
Avance	$99.276	$84.298	$114.297	$69.300	$159.295
Crédito consumo	$43.487	$23.498	$63.497	$3.501	$123.495
Total anual	$2.989.500	$2.875.114	$3.105.106	$2.760.129	$3.450.095
Total anual valor presente	**$2.989.500**	**$2.567.066**	**$2.475.371**	**$1.964.605**	**$2.192.598**

Tabla 43: Flujos de dinero implementación proyecto redes

A partir de lo anterior y considerando una inversión inicial de $3.200.000[22], se obtiene un valor actual neto (VAN) de **$8.989.140**.

[22] Inversión inicial calculada como el sueldo promedio por 2 meses para una ingeniero o ingeniero con las capacidades técnicas para la realización de un proyecto de esta naturaleza. Cabe destacar que no se consideran otros gastos para la implementación del proyecto debido a que los costos de procesamiento y obtención de datos se consideran marginales dado el contexto de la organización.

12 CONCLUSIONES

Las redes son una herramienta de gran impacto que permiten aprovechar y explotar datos donde no es posible obtener insight con ninguna otra metodología. La bibliografía muestra evidencia de que, incluso utilizando relaciones por reglas generales tales como zonas de residencia, profesión, hobbies, entre otros, es posible realizar análisis que permitan generar insights concluyentes.

En línea con lo anterior, abordar relaciones desde la base de transferencias del banco parece ser una estimación muy precisa de las relaciones que acontecen en la vida real y que se buscan representar en una red. Parece natural entonces pensar que, a partir de la estructura de transferencias, es posible identificar a las personas más relevantes para el banco. Sin embargo, con esta aseveración surge una pregunta fundamental. ¿Qué se considera una persona relevante?

En este proyecto se cuantifica la relevancia de un cliente como el potencial de apertura de tarjetas de crédito. Si bien existen infinitas definiciones de lo que es relevancia, este enfoque resulta adecuado a la hora de intentar contrastar los modelos aquí planteados con un hecho que sea cuantificable bajo los estándares y limitaciones de un proyecto de esta naturaleza. Esto último debe ser un factor relevante a la hora de mirar los análisis aquí realizados. El desarrollo de este proyecto comprende una metodología exhaustiva comenzando desde el procesamiento de grandes volúmenes de datos de muy baja calidad, hasta la evaluación estadística de métricas obtenidas para explicar el desempeño en una campaña a nivel de cliente. Por lo tanto, el potencial de esta metodología debe ser evaluado mirando el infinito abanico de posibilidades donde es posible aplicar estudios de esta naturaleza, los cuales van más allá del contexto de este tipo de campañas y de esta industria.

En cuanto al desarrollo del proyecto se considera que es posible validar 3 de las 4 hipótesis planteadas. En primer lugar, es directo inferir del proyecto que la base de transferencia corresponde a una buena estimación de las interacciones entre personas. Por consiguiente, hubiera sido posible aplicar solo modelos de machine learning a las características de las relaciones de clientes para obtener una aproximación de lo que sería el desempeño de los clientes en la campaña de referidos.

Respecto a la segunda hipótesis se concluye que las métricas obtenidas de los grafos si permiten distinguir entre clientes que, en promedio, tienen una mejor tasa de respuesta a campaña de referidos. Este efecto se pudo observar tanto para la campaña del 2018 como para el piloto de este proyecto, donde si bien, no todas las tasas de respuesta de los conjuntos eran comparables entre sí, en todos los modelos fue posible distinguir un conjunto con una tasa diferenciable y superior en promedio a la del resto de clientes.

La tercera hipótesis plantea que aquel cliente que se identifica como más relevante para el banco efectivamente lo sea. Esto también se considera probado ya que los clientes que se identifican como más relevantes efectivamente tienen, en promedio, una mayor tasa de respuesta a la campaña de referidos. Dado su potencial de apertura conocido, se puede asumir que tendrán un desempeño similar con incentivos más pequeños, pudiendo apalancar esfuerzo en ellos para realizar campañas con un desempeño similar a un menor costo.

La cuarta hipótesis plantea que es posible distinguir el efecto entre ambas métricas que describen la relevancia de un cliente. Esta es la única hipótesis que no se considera validada ya que ninguno de los modelos probados conto con la bondad de ajuste necesarias para mostrar con evidencia estadística el efecto de la dimensión nodo o enlace. Sin embargo, se considera que se encontraron claros indicios del sentido al que apunta este resultado ya que, si bien las variables analizadas no eran estadísticamente significativas, si entregaron resultados consistentes con lo que se observó en los análisis de los modelos no supervisados.

A partir de lo mencionado anteriormente, se identifican dos factores fundamentales que podrían haber perjudicado la calidad de los análisis: la baja tasa de respuesta del piloto o una baja calidad de las métricas estudiadas. Respecto a la tasa de respuesta se debe tener en consideración que la campaña se envió a un número acotado de clientes, lo que reduce la significancia estadística de los modelos estudiados. A su vez, existieron fuertes factores externos que sin duda afectaron el comportamiento habitual de las personas. Estos factores, marcados por importantes revueltas sociales, tuvieron un gran impacto en toda la industria financiera debido a un aumento de la incertidumbre política y económica en la sociedad.

Por otro lado, respecto a la calidad de las métricas estudiadas, se desconoce si fueron o no afectadas por la reducción del volumen de datos que finalmente se utilizó para los análisis. Si bien este factor no impide validar las otras hipótesis, se desconoce el real efecto que tuvo sobre la calidad de las métricas de la red.

Para finalizar, se presentó una metodología innovadora en la industria financiera utilizando algoritmos desarrollados y aplicados principalmente con objetivos académicos, con muy pocas aplicaciones en empresas. De esta forma, se logra obtener insights a partir de información que no había sido explotada anteriormente los que, a pesar de no ser concluyentes estadísticamente, si muestran un claro y creciente potencial el cual está directamente enraizado al mundo contemporáneo, cada vez más conectado y globalizado.

13 INVESTIGACIONES FUTURAS

A partir de los conocimientos adquiridos durante el desarrollo de este proyecto, es posible establecer recomendaciones de investigación que podrían ser de interés entendiendo la naturaleza del negocio y la factibilidad en cuanto al procesamiento de variables y análisis que ofrece la teoría de grafos.

Estas recomendaciones se pueden dividir en aquellas que podrían ser anexadas de forma directa a este trabajo, es decir, ser una continuación directa sobre el entregable de este proyecto. Por otro lado, también se plantean recomendaciones con un foco apartado al de este proyecto pero que podrían entregar insight de interés para el banco.

13.1 Relacionadas directamente al proyecto

Hubiera sido un aporte considerable el poder caracterizar los clústers encontrados en el análisis de la campaña de referidos del 2018. Esto habría dado una imagen más clara de que tan extrapolable era la relación que se observó en este análisis para ser utilizados como insight para la campaña del 2019. A la vez, habría permitido conocer las características en detalle de aquellos clientes pertenecientes a los conjuntos de mejor tasa de respuesta. A partir de allí surgen preguntas del tipo, ¿Las métricas de las redes (dimensión **nodos**) esta correlacionada con el comportamiento del cliente? ¿Qué el cliente pertenezca a un clúster hace que se comporte de cierta manera o es que, dado que se comporta de cierta manera, entonces tiene una mejor red?

Por otro lado, dado que la relación se obtiene de la matriz de transferencias, es posible realizar un análisis análogo al ya hecho, pero con grafos dirigidos, es decir, considerando la dirección y magnitud de las relaciones.

Por último, se recomienda el estudio de un piloto integrando variabilidad en los incentivos por aperturas. A partir del trabajo surgen interrogantes respecto a cómo las características de los clientes en cada clúster puedan afectar el comportamiento frente a una variación del incentivo de una campaña. En particular, se recomienda estudiar el efecto sobre la tasa de respuesta de los clientes pertenecientes al segmento más relevante frente a una baja en el incentivo por apertura.

13.2 Relacionadas al estudio de redes

Se recomienda generar nuevos estudios, contrastando resultados con el efecto de influencia de los clientes sobre sus conexiones bajo un enfoque distinto al aquí utilizado. Por ejemplo, modelos de gestión de reclamos. Para esto, podría ser interesante estudiar redes en distintos periodos de tiempo evaluando si aquellos clientes que se caracterizaron como relevantes, y que hayan presentado reclamos, han influenciado en el comportamiento de compra de los clientes con lo que se relacionan.

Otro enfoque donde se distingue un potencial a estos análisis es el estudio de relevancia a sucursales. Es posible realizar un análisis basado en la relación cliente-sucursal y obtener insights que ayuden a entender como una sucursal podría potenciar su gestión de espacios, productos o actividades de interés que los clientes puedan realizar en ella. Similar al enfoque anterior, se podría generar un modelo en que se analiza las relaciones de los clientes con el comercio (externo al holding) intentando cuantificar la relevancia de los distintos comercios con que se relacionan los clientes del banco, evaluar las características de ellos y a partir de allí tomar decisiones respecto a, por ejemplo, futuras alianzas.

Finalmente, es posible realizar un análisis bajo el enfoque de redes sobre el programa de fidelización del banco basado en puntos. Actualmente, este programa entrega beneficios diferenciados tanto en la acumulación como en el canje de acuerdo con la categoría de un cliente. Dicha categoría depende directamente de su nivel de consumo. Bajo el enfoque de redes, no considerar el poder de influencia de un cliente sobre las personas con las que interactúa genera una clasificación errónea, por lo tanto, se podría estar sub o sobre estimando la categoría de un cliente, perdiendo valor en la utilización del programa de fidelización.

14 BIBLIOGRÁFIA

[1]. Julia Klier, Mathias Klier, Florian Probst, Lea Thiel1 2014. Customer Lifetime Network Value. Fraunhofer.

[2]. Pedro Domingos, Matt Richardson. 2001. Mining the NEtowrk Value of Customers.

[3]. Gallardo M., C. E. 2016. Identificación de clientes con patrones de alta interacción con los drivers de una tarjeta de crédito. Memoria de Ingeniero Civil Industrial. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[4]. Melo G., J. P. 2017. Efecto en el comportamiento de compra al modificar los niveles de canje de un programa de fidelización. Memoria de Ingeniero Civil Industrial. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[5]. Aguirre SM., V. I. 2017. Perfilamiento de clientes influenciables en campañas de productos financieros en una empresa de retail financiero. Memoria de Ingeniero Civil Industrial. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[6]. Eight Leaves. Using RFM to identify your best consumers. Sirdhar Mutyala. 2017. [en línea] https://www.eightleaves.com/2011/01/using-rfm-to-identify-your-best-customers/ [consulta: 31 mayo 2019]

[7]. Parmenter, David (3 de abril de 2015). Key Performance Indicators: Developing, Implementing, and Using Winning KPIs (3ra edición). John Wiley & Sons. p. 99.

[8]. GeeksforGeeks. Page rank algorithm and implementation. Anamitra Musib 2015. [en línea] https://www.geeksforgeeks.org/page-rank-algorithm-implementation/ [consulta: 31 mayo 2019]

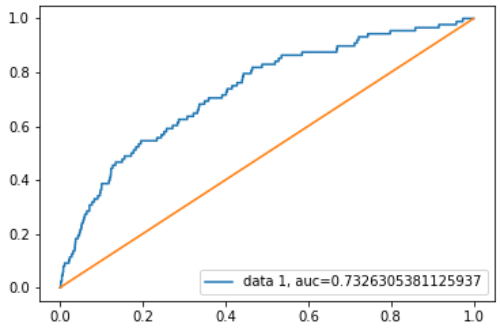
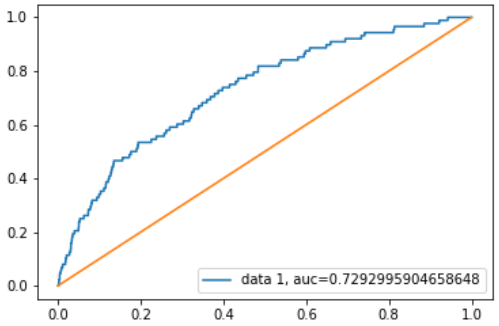
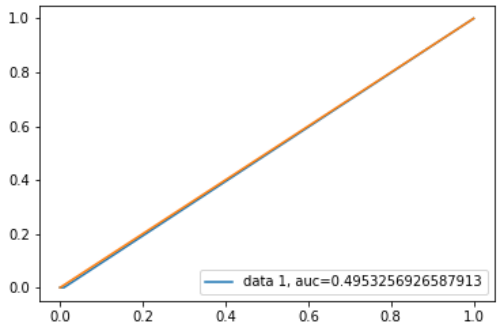
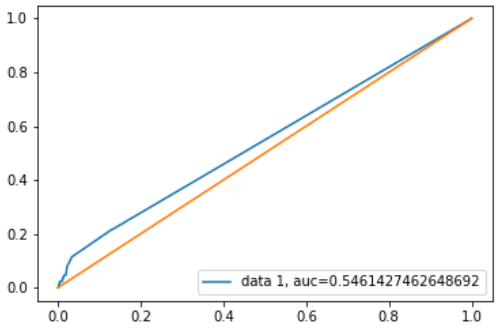
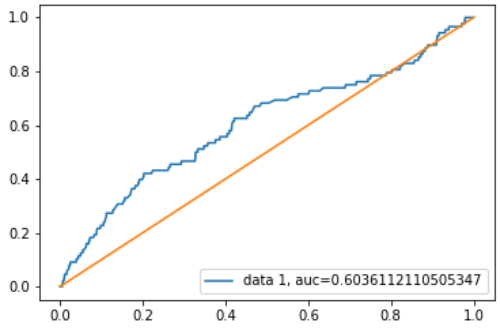
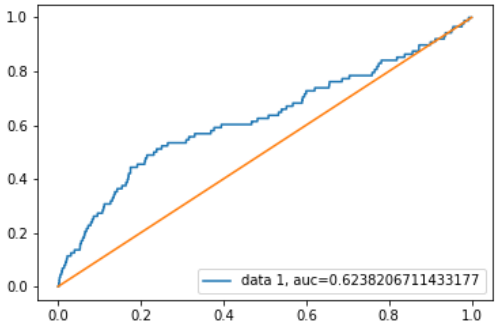
[9]. Techopedia. Knowledge discovery in databases (KDD). 2017. [en línea] https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd [consulta: 31 mayo 2019]

[10]. Memoria Anual 2019 Banco Cliente

[11]. Documento Integracion Banco Cliente

15 ANEXOS

15.1 Anexo sección 9.1.5

1.1 Curva ROC Logistic Regression	1.2 Curva ROC Logistic Regression Balanced
 *Ilustración 81: Curva ROC Modelo Logistic Regression*	 *Ilustración 82: Curva ROC Modelo Logistic regression balanced*
2.1 Curva ROC Random Forest	2.2 Curva ROC Random Forest Balanced
 *Ilustración 83: Curva ROC Modelo Random Forest*	 *Ilustración 84: Curva ROC Modelo Random Forest balanced*
3.1 Curva ROC Gradient Boosting	3.2 Curva ROC Gradient Boosting Balanced
 *Ilustración 85: Curva ROC Modelo Gradient Boosting*	 *Ilustración 86: Curva ROC Modelo Gradient Boosting Balanced*
4.1 Curva ROC Ada Boost	4.2 Curva ROC Ada Boost Balanced

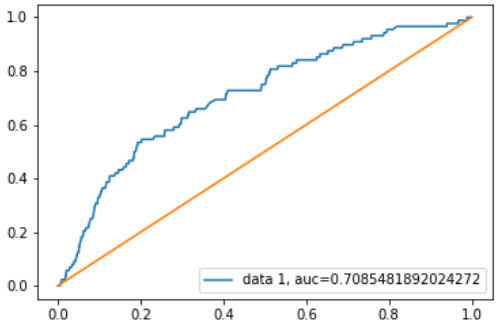


Ilustración 87:Curva ROC Modelo Ada Boost

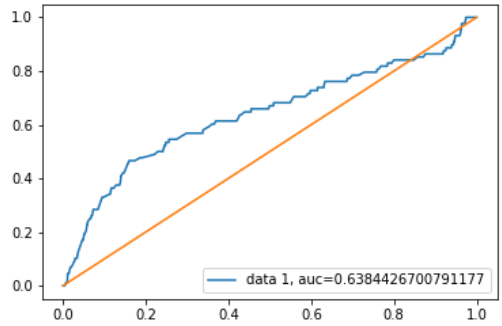


Ilustración 88: Curva ROC Modelo Ada Boost Balanced

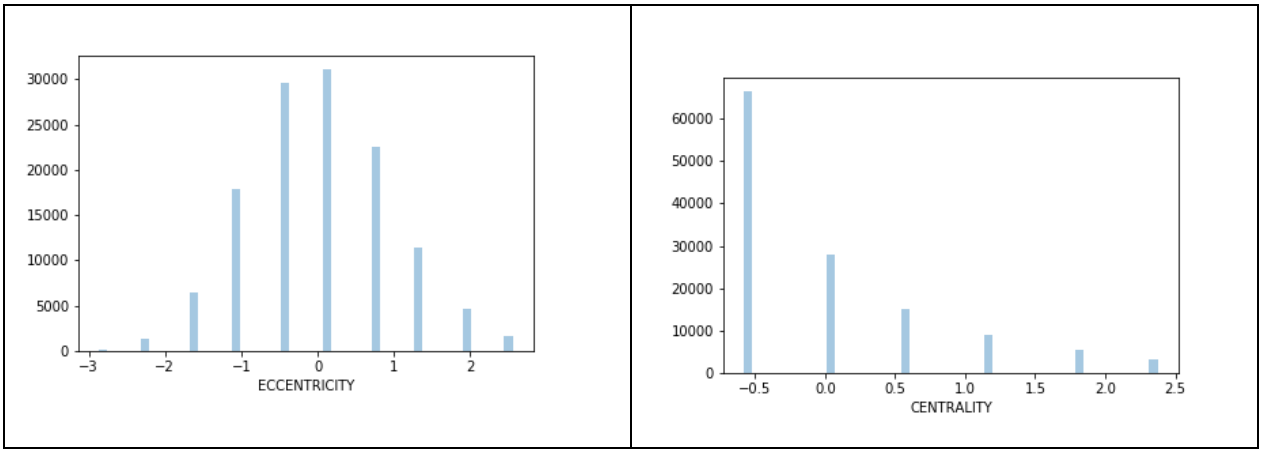
15.2 Anexo sección 9.1.6.1

Iteración	Nodos	Enlaces	Numero relaciones eliminadas	Nodos eliminados	Nodos Refieren
0	1153059	1187513	-	-	1154
1	323460	424423	1	799471	691
2	258810	364757	1	62866	577
3	233581	340224	1	24991	550
4	225885	332626	1	7666	537
5	222487	329248	1	3389	533
6	221257	328018	1	1230	532
7	220692	327453	1	565	532
8	220448	327209	1	244	532
9	220351	327112	1	97	532
10	220312	327073	1	39	532
11	220300	327061	1	12	532
12	220294	327055	1	6	532
13	220292	327053	1	2	532
14	217791	324434	1	0	515

15.3 Anexos sección 9.1.9

Las variables Eccentricity y Centrality son métricas representadas por números enteros. Esto provoca que, si se estudia su distribución, observen valores concentrada en ciertos puntos. Tal como se puede observar en las siguientes gráficas:

ECCENTRICITY	CENTRALITY



15.4 Anexos sección 9.2.1

1. Monto

	MONTO
count	34.346.408
mean	103.714
std	297.571
min	2
25%	10.000
50%	29.000
75%	100.000
max	180.000.000

Tabla 44: Descripción monto transferencias 2018/2019

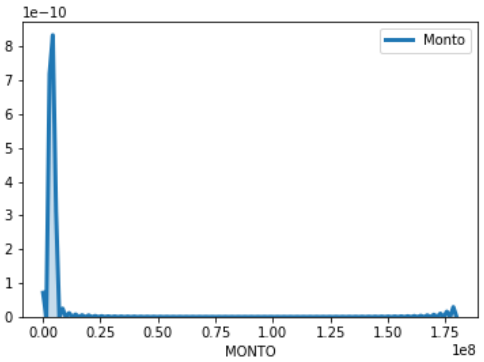


Ilustración 89: Distribución monto transferencias 2018/2019

2. Relaciones

	Relaciones
count	8.122.389
mean	4
std	10
min	1
25%	1
50%	1
75%	3
max	2.629

Tabla 45: Descripción relaciones únicas transferencias 2018/2019

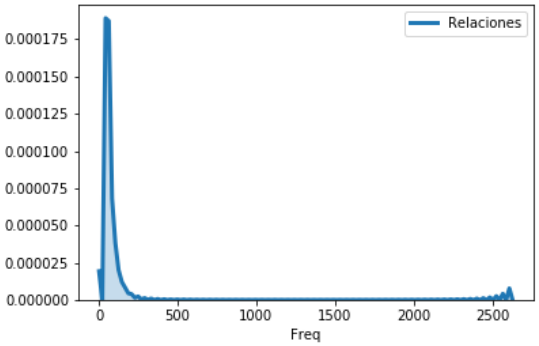


Ilustración 90: Distribución relaciones únicas transferencias 2018/2019

15.5 Anexos 9.2.2.1

1.1 Distribución variable Delta:	1.2 Boxplot variable Delta:

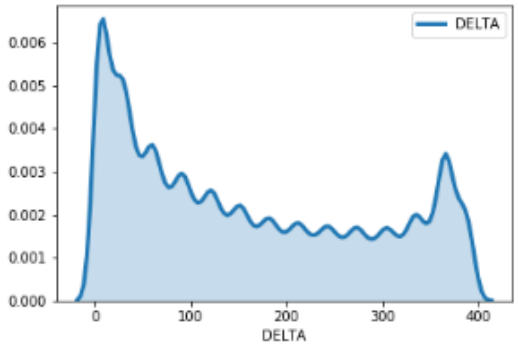
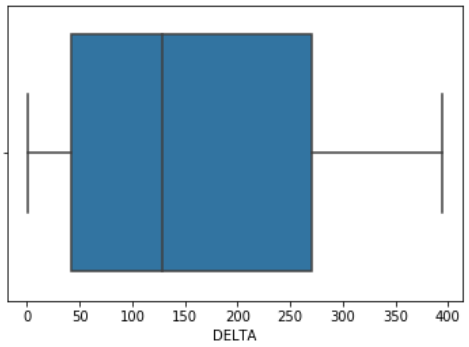

	

Ilustración 91: Distribución variable Delta	*Ilustración 92: Boxplot variable Delta*
2.1 Distribución variable Frequency:	**2.2 Boxplot variable Frequency:**

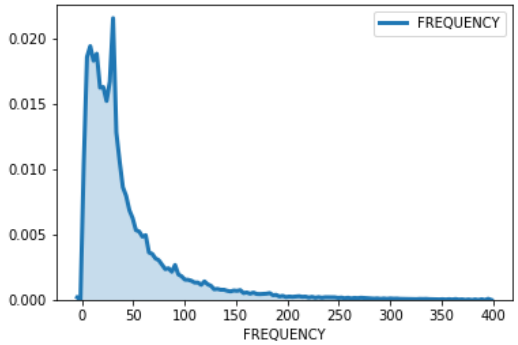
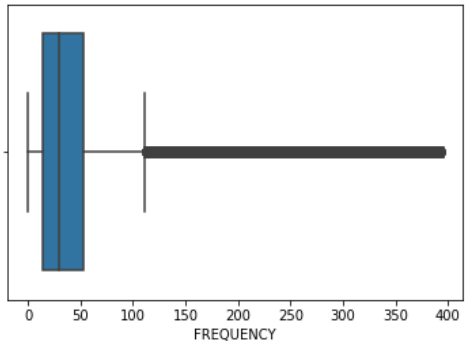

	

Ilustración 93: Distribución variable Frequency	*Ilustración 94: Boxplot variable Frequency*
3.1 Distribución variable Monto	**3.2 Boxplot variable Monto**

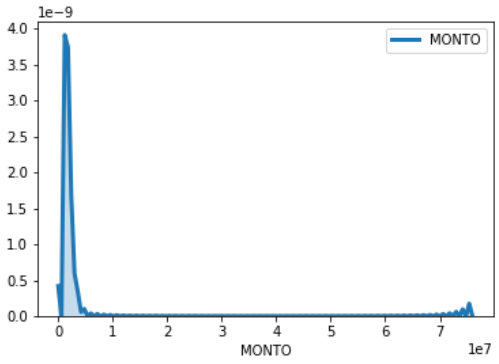
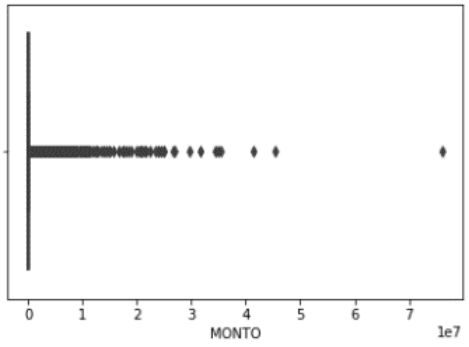

	

Ilustración 95: Distribución variable Monto	*Ilustración 96: Boxplot variable Monto*
4.1 Distribución variable Recency	**4.2 Boxplot variable Recency**

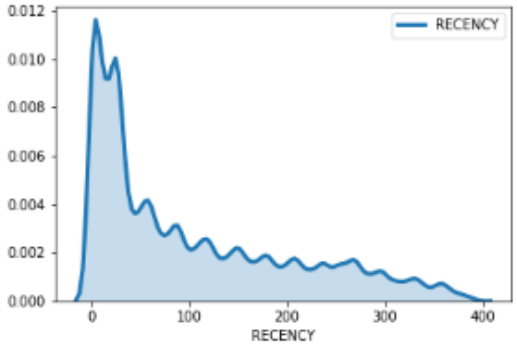


Ilustración 97: Distribución variable Recency

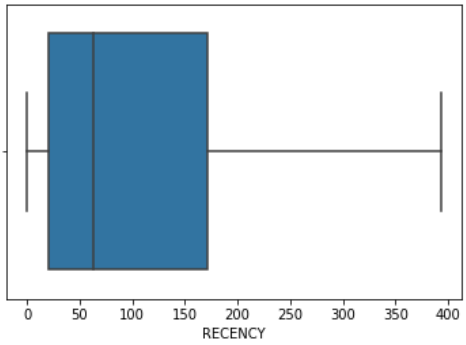


Ilustración 98: Boxplot variable Recency

5.1 Distribución variable RF:

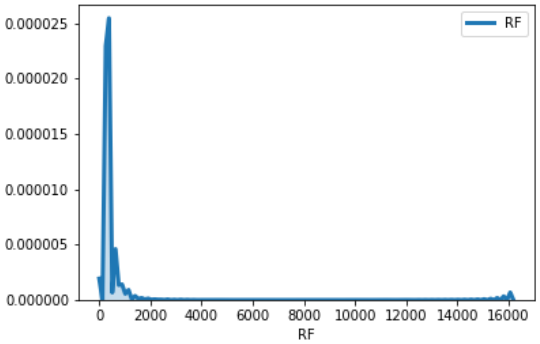


Ilustración 99: Distribución variable RF

5.2 Boxplot variable RF:

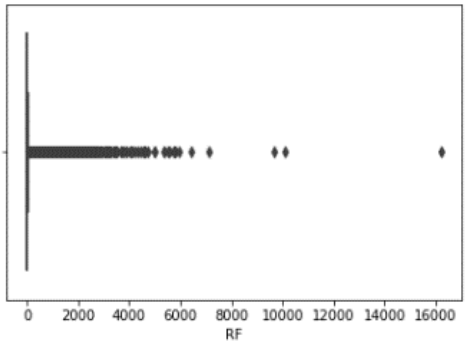


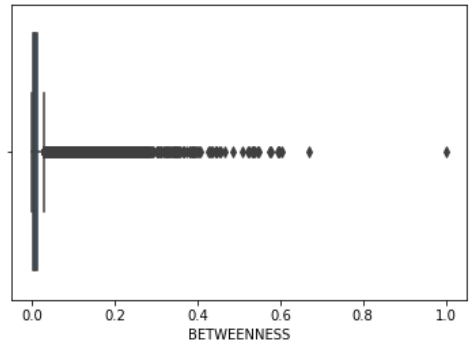
Ilustración 100: Boxplot variable RF

15.6 Anexo sección 9.2.4.1

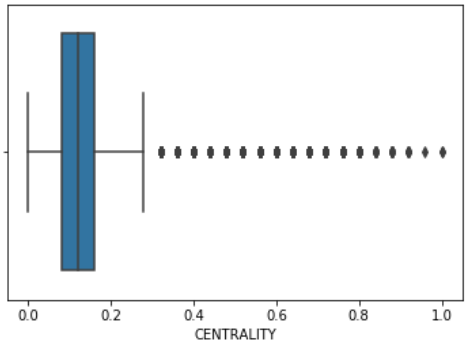
Iteración	Nodos	Enlaces	Numero relaciones eliminadas	Nodos eliminados
0	2359883	2961343	0	0
1	823893	1475049	1	1535990
2	762045	1414843	1	61848
3	740034	1392913	1	22011
4	736936	1389820	1	3098
5	735931	1388815	1	1005
6	735761	1388645	1	170
7	735712	1388596	1	49
8	735703	1388587	1	9
9	735698	1388582	1	5
10	735696	1388580	1	2
11	735695	1388579	1	1
12	364305	708665	2	371390
13	335856	680840	1	28449

15.7 Anexo sección 9.2.5

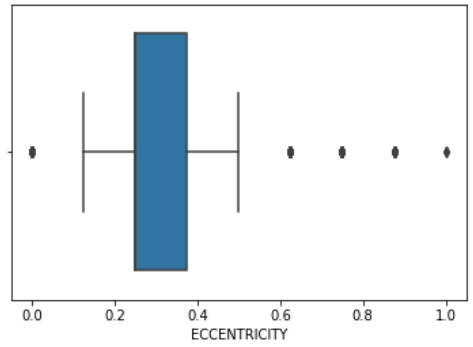
1. Boxplot Betweenness



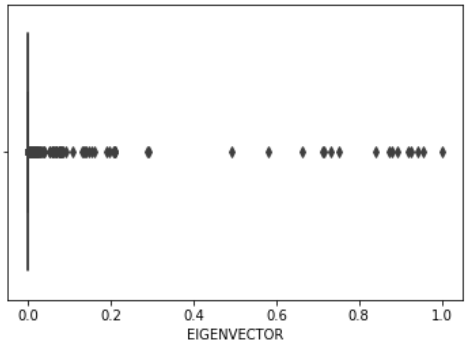
2. Boxplot Centrality



3. Boxplot Eccentricity



4. Boxplot Eigenvector



15.8 Anexo sección 11

Supuestos justificación económica.

Producto	Rentabilidad promedio anual	Porcentaje	Universo clientes	Porcentaje clientes contactables reales	Porcentaje clientes relevantes
Tarjeta crédito	$50.000	85%	42.500	13%	25%
Super avance	$30.000	12%	6.000	13%	25%
Avance	$15.000	12%	6.000	13%	25%
Crédito consumo	$20.000	5%	2.500	13%	25%

Flujos aperturas campaña año 1.

Año 1 campaña normal			Año 1 campaña clientes relevantes		
Universo clientes	Tasa respuesta	Aperturas	Universo clientes	Tasa respuesta	Aperturas
371.875	0,0007299	272	92.969	0,0035	327
52.500	0,0007299	39	13.125	0,0035	48
52.500	0,0007299	39	13.125	0,0035	48
21.875	0,0007299	17	5.469	0,0035	21

Flujos aperturas campaña año 2.

Año 2 campaña normal			Año 2 campaña clientes relevantes		
Universo clientes	Tasa respuesta	Aperturas	Universo clientes	Tasa respuesta	Aperturas
371.909	0,0007299	271	92.979	0,0035	327
52.505	0,0007299	38	13.126	0,0035	48
52.505	0,0007299	38	13.126	0,0035	48
21.877	0,0007299	16	5.469	0,0035	21

Flujos aperturas campaña año 3.

Año 3 campaña normal			Año 3 campaña clientes relevantes		
Universo clientes	Tasa respuesta	Aperturas	Universo clientes	Tasa respuesta	Aperturas
371.909	0,0007299	272	92.979	0,0035	326
52.505	0,0007299	39	13.126	0,0035	47
52.505	0,0007299	39	13.126	0,0035	47
21.877	0,0007299	17	5.469	0,0035	20

Flujos aperturas campaña año 4.

Año 4 campaña normal			Año 4 campaña clientes relevantes		
Universo clientes	Tasa respuesta	Aperturas	Universo clientes	Tasa respuesta	Aperturas
371.909	0,00072989	272	92.979	0,0035	327
52.505	0,00072989	39	13.126	0,0035	48
52.505	0,00072989	39	13.126	0,0035	48
21.877	0,00072989	17	5.469	0,0035	21

Flujos aperturas campaña año 5.

Año 5 campaña normal			Año 5 campaña clientes relevantes		
Universo clientes	**Tasa respuesta**	**Aperturas**	**Universo clientes**	**Tasa respuesta**	**Aperturas**
371.909	0,00072989	278	92.979	0,0035	326
52.505	0,00072989	45	13.126	0,0035	47
52.505	0,00072989	45	13.126	0,0035	47
21.877	0,00072989	23	5.469	0,0035	20

Diferencia aperturas por año.

	Año 1	**Año 2**	**Año 3**	**Año 4**	**Año 5**	**Total**
	Delta	**Delta**	**Delta**	**Delta**	**Delta**	**Delta**
Tarjeta crédito	54	52	56	57	56	275
Super avance	8	6	10	11	10	43
Avance	8	6	10	11	10	43
Crédito consumo	3	1	5	6	5	21

Valor presente neto proyecto.

Producto	Año 1	Año 2	Año 3	Año 4	Año 5
Tarjeta crédito	$2.648.185	$2.598.722	$2.698.719	$2.548.729	$2.848.714
Super avance	$198.552	$168.596	$228.594	$138.600	$318.591
Avance	$99.276	$84.298	$114.297	$69.300	$159.295
Crédito consumo	$43.487	$23.498	$63.497	$3.501	$123.495
Total anual	$2.989.500	$2.875.114	$3.105.106	$2.760.129	$3.450.095
Total anual valor presente	$2.989.500	$2.567.066	$2.475.371	$1.964.605	$2.192.598
Inversión	$3.200.000				
VAN	**$8.989.140**				