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

DISEÑO DE UN MODELO DE RECOMENDACIONES DE NEXT BEST ACTION CORPORATIVO EN LA LÍNEA RETAIL DE UN HOLDING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ANDRÉS IGNACIO CONTRERAS TAPIA

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
CAROLINA SEGOVIA RIQUELME
ANDRES MUSALEM SAID

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Andrés Ignacio Contreras Tapia
FECHA: 21/11/2019
PROFESOR GUÍA: Pablo Marín Vicuña

DISEÑO DE UN MODELO DE RECOMENDACIONES DE NEXT BEST ACTION CORPORATIVO EN LA LÍNEA RETAIL DE UN HOLDING

En esta memoria se plantea la gestión del valor de los clientes con perspectiva de rentabilidad mediante hitos comerciales. La rentabilidad de un hito se estima como el efecto causal en gasto, contribución bruta y transacciones que generó en promedio la realización de los hitos en el pasado. La memoria tiene un enfoque corporativo, por lo que se incorpora información de 3 negocios de retail: Formato 1 (F1), Formato 2 (F2) y Formato 3 (F3). Los hitos estudiados son: Primera compra web, apertura de tarjeta de crédito, Cruce de negocio F1 a F2 y cruce de negocio F2 a F1.

Se utiliza la metodología de Propensity Score Matching (PSM), que permite estimar efectos causales en estudios observacionales. El problema del contrafactual de los clientes activados (cómo habría sido el gasto del cliente en ausencia del hito) se resuelve seleccionando clientes comparables del grupo que no se activó en el pasado. Con la finalidad de obtener efectos heterogéneos se realizan segmentaciones separando a los clientes en grupos de propensión.

Así, la primera parte de la metodología consiste en estimar los puntajes de propensión de los hitos en enero 2017. Para ello, se prueban 3 modelos de clasificación distintos: Regresión logística, Random Forest y Red Neuronal. El modelo se selecciona en base a la calidad de emparejamiento logrado, siendo seleccionada en la mayoría de los casos la Red Neuronal y en 1 caso la Regresión Logística. Una vez hecho estos pasos anteriores se procede a calcular los efectos causales mediante diferencia en diferencia de los clientes activados y los no activados comparables.

Se concluye que todos los hitos representaron un efecto causal positivo. El efecto más grande en gasto es de la apertura de tarjeta que representa un aumento de 260%, en segundo y tercer lugar, están empatados el paso del Formato 1 al 2 y del 2 al 1 con 47% de incremento de gasto, y en el último lugar está la primera compra web con un incremento de 32%. En contribución, el primero es la tarjeta con 170%, el segundo el paso de F1 a F2 con 58%, tercero el paso de F2 a F1 con 47% y nuevamente el último es la primera compra web con un 27%. El valor agregado de gestionar a los clientes con estos hitos representa una oportunidad importante pues existen muchos clientes que aún no los hacen. Estos resultados se mostraron robustos ante variaciones la metodología. Por el lado de los resultados en tarjeta un caso interesante sería probar los resultados, pero agregando información de rentabilidad de los negocios financieros. Por otro lado, el valor agregado de la gestión debe considerar el costo en descuentos o cupones que se puedan utilizar para desencadenar los hitos.

Finalmente, la metodología se mostró útil para encontrar el valor de los hitos. Para pasar a una fase de producción, la metodología se puede complementar, ampliando el estudio para considerar más aspectos relevantes, como cambiar la ventana de evaluación de los efectos causales y desagregar la identificación de los hitos (Ejemplo: Primera compra web separado en cada negocio vs en el holding) Una de las propuestas es analizar cómo cambian los resultados de la memoria utilizando experimentación y en contexto de email marketing, viendo como varía el gasto de los clientes que se activan al participar de alguna campaña.

1. TABLA DE CONTENIDO

2. INTRODUCCIÓN

2.1 CONTECTO EMRPESARIAL

El contexto empresarial de la corporación es el mercado del retail chileno y latinoamericano, pero toda la memoria se analizará desde la perspectiva de Chile.

El mercado chileno del retail está presentando tasas de crecimiento por debajo de los niveles históricos. Hoy se prevé que el crecimiento real anual del sector se ubicará en torno a los 2,1% hasta el año 2021 (Emol, 2017), a diferencia de la alta tasa de crecimiento que el sector promedió durante los años 2011 – 2016. Esto es una señal de la madurez que ha logrado el mercado chileno, donde las estrategias están enfocándose en sacar mejor provecho a la base de clientes de las empresas y mejorar los procesos en los canales de e-commerce.

En el gráfico 1 se puede observar que el crecimiento del comercio a nivel nacional solo ha estado en peores condiciones que las actuales en 3 ocasiones, 2 de ellas durante crisis mundiales. Las fuentes anteriormente nombradas muestran que los crecimientos de la industria en Chile se encuentran en mínimos históricos y esto se debe a la madurez del mercado, por lo que a diferencia del pasado las expectativas de una recuperación de esta métrica son menores.

Gráfico 1: crecimiento de comercio minorista en Chile, CNC 2018

Frente a las menores perspectivas de crecimiento que enfrenta este sector de la economía chilena es que han nacido nuevas tendencias de la industria que estarán liderando los prospectos de desarrollo de las empresas de este tipo.
En America Retail se resume que las tendencias del retail que marcarán la agenda del futuro según Accenture serán la inteligencia artificial, la transformación digital enfocando las soluciones digitales a las necesidades humanas. Por otro lado, se promueve el funcionamiento de la compañía como ecosistema, donde no existan barreras entre los canales físicos y webs dado que los clientes esperan que los canales estén coordinados, funcionen de forma conjunta y entreguen valor a los usuarios de una forma integral (America Retail, diciembre 2017).

2.2 EMPRESA

La empresa en la que se desarrollará la memoria corresponde a un holding del retail que tiene presencia en los rubros de Formato 1, Formato 2 y Formato 3. Además, cuenta con una línea de negocios del tipo financiero, donde una de las actividades más destacadas es la administración de tarjetas de crédito del tipo casa comercial. La tarjeta funciona como vínculo transversal de la corporación entregando ofertas exclusivas en todos los retail siendo el vehículo del sistema de fidelización de la corporación, mediante campañas y programas de acumulación de puntos.

La empresa durante los años 2011-2015 mostró crecimientos anuales en ventas por sobre los dos dígitos (gráfico 2), pero este crecimiento se vio abruptamente reducido desde el año 2015 en adelante. Esto se condice con el crecimiento promedio anual de las ventas del sector de retail durante los mismos años y su posterior reducción. De todas formas, la corporación promedió crecimientos anuales de un 13% entre el 2011-2015, lo que significa que creció más que lo que creció el mercado del retail en Chile en ese periodo.

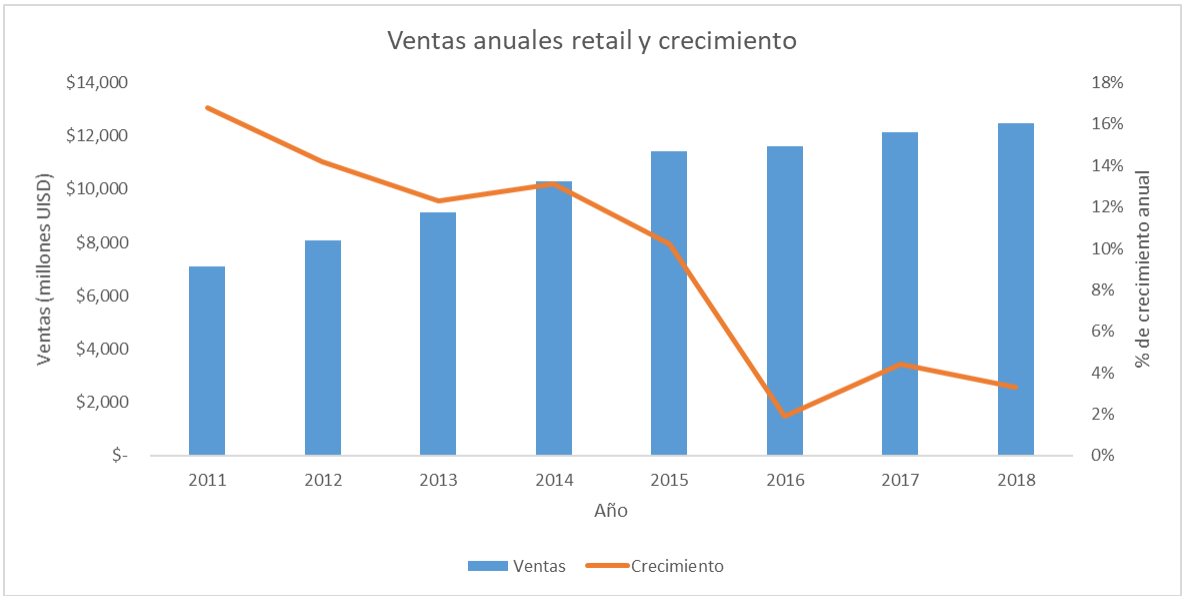


Gráfico 2: Ventas y crecimiento anual Retail (Análisis razonado 2011-2018, referencia oculta)

En el gráfico 3 se hace un análisis más detallado de la cantidad de clientes por año del holding, donde se puede apreciar cómo han evolucionado a medida que avanza el tiempo, en cada canal. Tanto en el canal físico como web se produce un aumento considerable de la cantidad de clientes por año

Gráfico 3: Clientes por negocio (Elaboración propia)

A continuación, se realiza un análisis del comportamiento de las ventas de los negocios a lo largo de los meses de los últimos años.

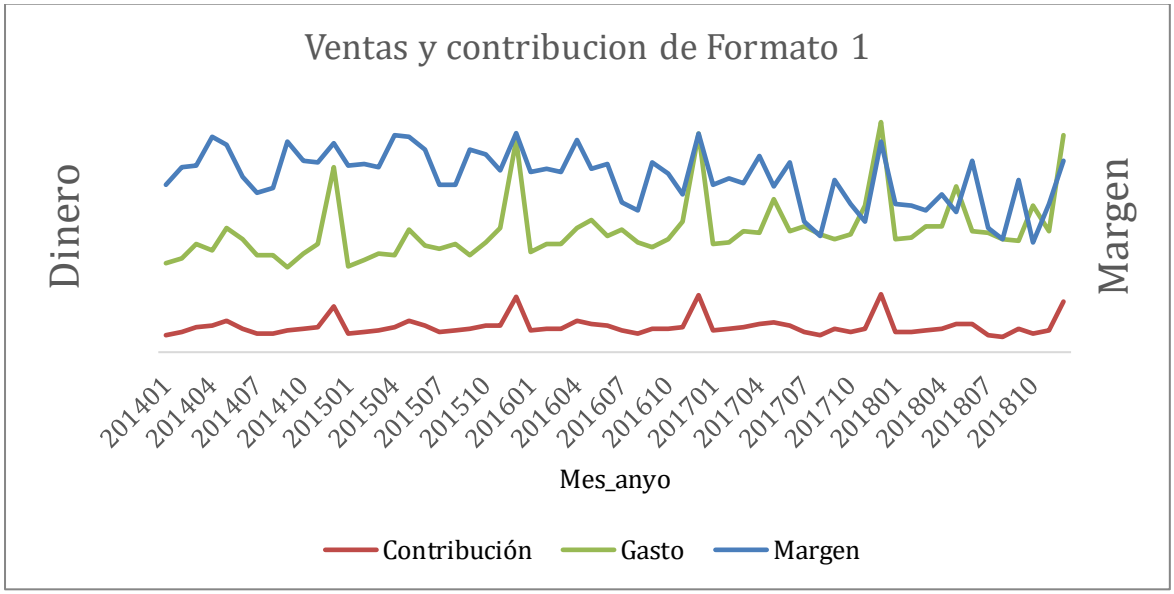


Gráfico 5: Contribución, gasto y margen para Formato 1 (Elaboración propia)

En el gráfico 5 se puede apreciar el alza de ventas del negocio Formato 1 desde el año 2014, pero a pesar de presentar esta alza, no sucede lo mismo para el caso de la contribución bruta por producto. La razón de porque sucede esto el margen bruto promedio por cliente ha ido disminuyendo con el tiempo. Por lo que se concluye que el negocio está vendiendo más, pero esta al mismo tiempo ha bajado la contribución bruta total.

Por el lado intra – año se aprecia que existe una estacionalidad clara en los meses de diciembre de cada año, donde se vende del orden de del doble que el resto del año promediado sin diciembre. También se observa una estacionalidad en los meses de mayo que se debe al efecto del CyberDay del primer semestre, el que se va acentuando con el tiempo lo que deja entrever que estas fechas han cobrado mayor relevancia para la empresa en los últimos años. Por otro lado, también está apareciendo cada vez más una estacionalidad entre octubre y noviembre que vendría a coincidir con el CyberDay del segundo semestre.

Un aspecto para destacar de las ventas y contribución mensual de Formato 1 es que se ve claramente la estacionalidad en la contribución para el caso de diciembre, pero no así para el resto de los meses. Por ejemplo, mayo y noviembre a pesar de tener un peak de ventas, esto no se traduce a contribución, lo que a priori se podría deber a que aumentan la cantidad de ofertas que se hacen durante esas fechas, bajando el margen bruto promedio.

En el caso de las tiendas de Formato 2 sucede algo distinto, pues existió una tendencia al alza durante el año 2017 de la venta identificada, principalmente por el hecho de que se está identificando de una mejor forma a los clientes. Esto gracias campañas realizadas para mejorar la identificación de los clientes durante 2015 y 2016. Por lo tanto, sería injusto comparar las tendencias de Formato 1 y Formato 2. Lo que sí se puede comprar es el nivel de margen de ambos negocios, donde se observa que, a diferencia del caso de Formato 1, este negocio no ha disminuido su margen bruto promedio con el paso de los años.

Otro punto relevante por mencionar es que, a diferencia del caso de Formato 1, este negocio no presenta estacionalidades perceptibles a la simple vista en ninguna de las 3 métricas

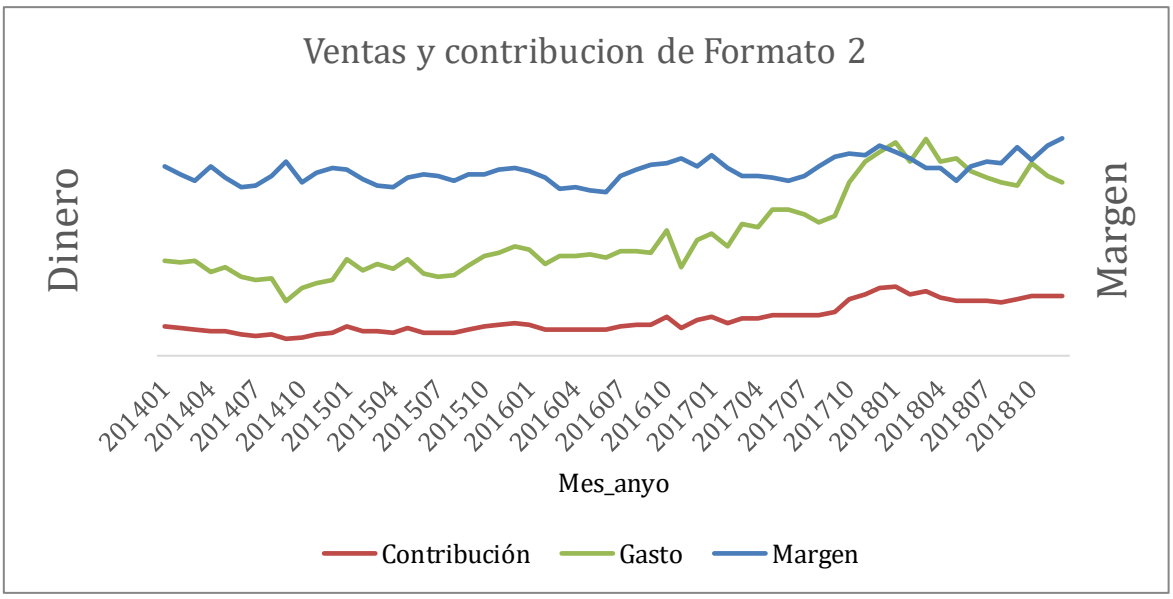


Gráfico 6: Contribución, gasto y margen para tiendas de Formato 2 (Elaboración propia)

Finalmente, el último negocio a analizar es el de Formato 3 (Gráfico 7). El negocio muestra niveles estables de venta y contribución y niveles de margen con una pequeña tendencia a la baja bordeando, por debajo de los otros negocios.

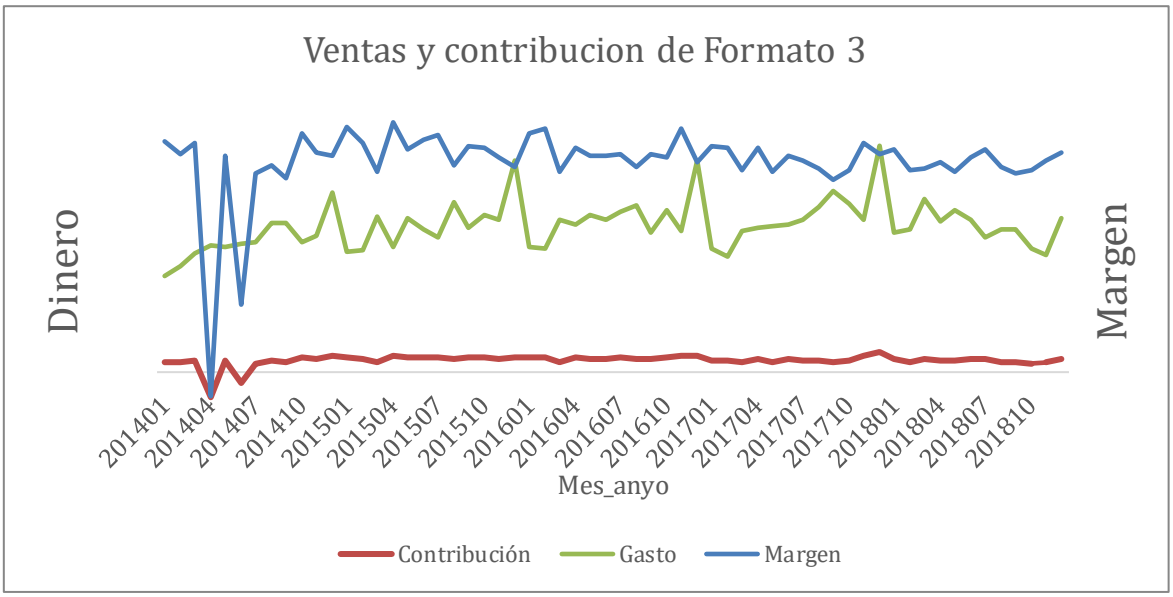


Gráfico 7: Contribución, gasto y margen para de Formato 3 (Elaboración propia)

2.3 DESCRIPCION DEL PROYECTO

Durante los últimos años, en el holding se ha segmentado el valor de los clientes solamente desde la perspectiva de cada negocio en particular y no desde el análisis global. Esto también sucede para la gestión de los clientes, donde cada negocio evalúa la rentabilidad de las distintas gestiones solamente considerando la perspectiva de sus ingresos particulares.

En este contexto es que resulta importante empezar a pensar en el cliente con una perspectiva global corporativa, considerando que los ingresos que percibe la corporación aportan al valor de la compañía de forma conjunta.

Así es relevante entender que, durante el ciclo de vida del cliente, éste se puede relacionar con el holding a través de cualquiera de sus unidades de negocio. Por lo que, dependiendo de las necesidades del cliente, la relación con un negocio en particular podría generar efectos cruzados a nivel corporativo, posiblemente generando aumentos de rentabilidad en uno y disminuciones en otro.

Un ejemplo de esto son los negocios de Formato 1 y Formato 3 que tienen categorías en común, por lo que si un cliente de un negocio cruza al otro se podría producir una canibalización de la venta. Desde la perspectiva individual de uno de los 2 negocios esto podría ser malo, pero desde la perspectiva global de la compañía esto podría ser bueno

pues el cliente podría empezar a comprar en nuevas categorías que el otro negocio no tenía disponible.

Otro ejemplo que muestra la relevancia de la perspectiva corporativa es el caso de la tarjeta comercial o de crédito del holding. El holding cuenta con la tarjeta como su principal elemento de fidelización y sirve de vehículo para distintas campañas y descuentos en todos los negocios. Por lo que el efecto que tiene la apertura de este servicio afecta de forma positiva a muchos negocios y no solamente al negocio que la administra.

Así es que frente a la disminución del crecimiento del retail en el país durante los últimos años y considerando que las perspectivas de crecimiento de estos rubros a nivel nacional se mantienen a la baja, dado el nivel de madurez (El mercurio, 2017) es que vuelve relevante aprovechar la oportunidad de la gestión global de los clientes del holding con miras a la rentabilidad total percibida al mismo tiempo que las unidades evalúan sus desempeños y sus oportunidades de aumento de valor de forma individual.

El proyecto busca ser un acercamiento a la oportunidad gestionar a los clientes de forma corporativa, utilizando como medida de rentabilidad los ingresos y contribuciones brutas percibidas por los negocios de Formato 1, Formato 2 y Supermercado.

En síntesis, se busca generar una recomendación de Next Best Action de gestión para cada cliente del holding, donde las recomendaciones serán hitos que correspondan a eventos de relación cliente-holding que se cree, a priori, puedan tener un efecto positivo en la rentabilidad del cliente dentro de la corporación.

Por lo que se toma de relación cliente-Holding se genera una herramienta que recomiende a la unidad corporativa de la empresa cuáles de estos hitos maximiza la utilidad esperada por cada cliente, teniendo en consideración la historia transaccional del mismo, las características sociodemográficas y la tenencia de productos financieros del holding al momento de la realización de la recomendación.

2.4 JUSTIFICACIÓN DEL PROYECTO

La justificación de este proyecto se presentará de dos formas. Primero se presentará la subclasificación del valor de los clientes hoy hecha por el holding. Esto con el objetivo de mostrar cómo podría mejorar la clasificación con solamente considerar la totalidad de los gastos y rentabilidades hechas por los clientes en él, con los 3 retail en forma conjunta.
Por otro lado, se mostrará la oportunidad que representa la gestión de 2 hitos de relación cliente-holding en términos de contribución y venta incremental, considerando cierto porcentaje de clientes gestionados satisfactoriamente.

2.4.1 GESTIÓN CORPORATIVA

Como fue mencionado anteriormente, la clasificación de valor de los clientes en las líneas del retail del Holding se hace de forma separada solamente evaluando los gastos de los clientes en cada formato de forma aislada.

Tipo de Cliente	Cantidad de Clientes
Alto Valor F1	~500,000
Alto Valor F2	~500,000
Alto Valor F3	~60,000

Tabla 1: Cantidad de clientes dependiendo del tipo de identificación de valor

En la tabla 1 se muestra que existe una suma importante de clientes que se consideran de alto valor en alguno de los 3 formatos del caso. Por otro lado, hoy existen clientes que no son considerados como alto valor por ningún negocio pero que podrían gastar una cantidad importante si es que se analizan desde la perspectiva corporativa.

Para lograr probar la existencia de estos se hará mediante la siguiente pregunta, si es que ordenamos el gasto de los clientes de menor a mayor, qué cantidad de clientes no identificados como valor en los negocios supera al 50% de los clientes que si son identificados como valor. Así en la tabla 2, se puede apreciar la cantidad de clientes que superan en gasto a la mediana de los clientes que son considerados como alto valor por los negocios. Las cifras han sido redondeadas, pero aun así se logra apreciar la cantidad de clientes que hoy no son considerados como alto valor, pero aun así gastan una cifra considerable.

Para probar que la visión corporativa podría complementar a la visión local de cada formato, se muestra en la tabla 2 la cantidad de clientes que no son considerados como alto valor en ningún formato pero que superan en gasto total a la mitad de los clientes que si son considerados como alto valor en cada negocio.

Tipo de Cliente	Superados por
Alto Valor F1	50,000
Alto Valor F2	25,000
Alto Valor F3	75,000

Tabla 2: Sub clasificación del valor

Así en la tabla 2 se muestra que 50 mil clientes que no son considerados como alto valor por ningún formato, superan al 50% de los clientes de alto valor formato 1, 25 mil clientes superan al 50% de alto valor del formato 2 y 75 mil clientes superan el gasto del 50% del alto valor del formato 3. Por lo tanto, se logra entender que la visión corporativa consideraría como alto valor clientes que en el pasado han sido pasados por alto según estas clasificaciones.

A continuación, veremos cómo distribuyen las contribuciones de los clientes que son detectados como valor por los negocios y aquellos que no lo son.

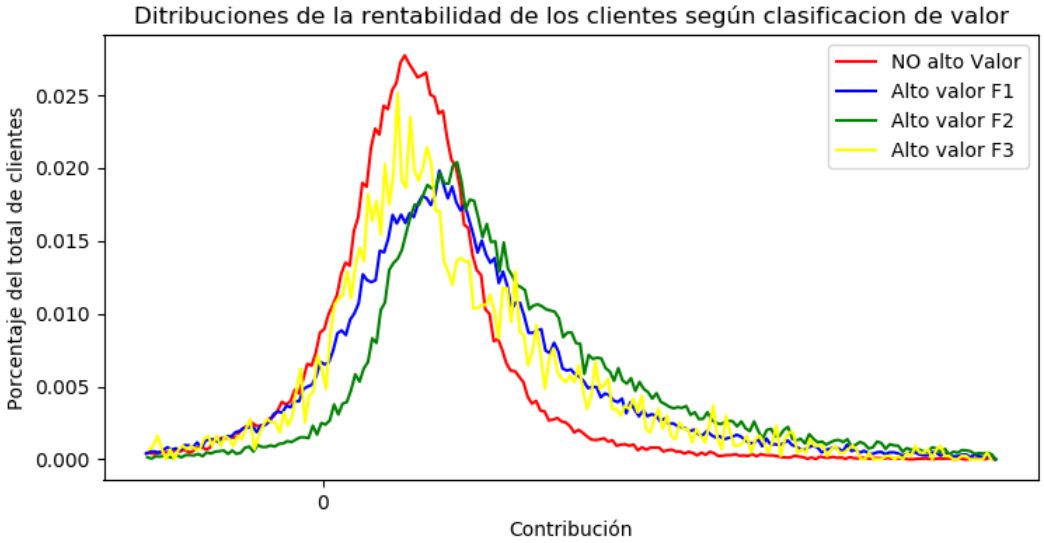


Gráfico 8: Contribución de los clientes según clasificación de valor

En el gráfico 8 se muestra a la distribución de rentabilidad (contribución bruta) por cliente condicionando a que el gasto sea superior a una cifra arbitraria y grande. Las distribuciones de rentabilidad para las personas que tienen más de esta cifra de gasto durante el 2018 en el holding son efectivamente distintas, en general los clientes que son considerados como valor han rentado más que los que no lo son. Esto se produce principalmente porque los clientes detectados como alto valor en algún negocio también gastan más a nivel corporativo que aquellos que no.

Por lo que tomando en consideración en cuenta el contexto del mercado del retail y los peores desempeños corporativos de los últimos años en comparación a lo observado en los años 2011 y 2015 es que se podría justificar una visión complementaria del valor local de los clientes y empezar una gestión corporativa que maneje el valor de los clientes con una visión global. A continuación, se mostrará la oportunidad de gestionar con 2 hitos a modo de ejemplo.

2.4.2 OPORTUNIDAD DE GESTIÓN

Para cuantificar el valor de la oportunidad de usar una herramienta que recomiende un nuevo tipo de gestión (la corporativa), es que se elegirán 2 hitos para evaluar el cambio en valor que se podría percibir por que el Holding realice las gestiones recomendadas.

El procedimiento para evaluar la oportunidad de gestión es:

1) Tomar clientes que durante el 2016 no realizaron el hito a estudiar

2) Durante el 2017 este grupo de personas se divide en 2: los que realizan el hito y los que no lo realizan.

3) Durante el año 2018 se ve cuánto es su gasto y rentabilidad a nivel Holding para poder compararlo con estas mismas métricas durante el 2016.

Los hitos para estudiar en la justificación serán:

- Abrir tarjeta comercial
- Comprar por primera vez en canal web

Para fines prácticos este análisis descriptivo se realiza con el 30% de los datos, lo que no representa una definición de alcance.

2.4.3 ABRIR TARJETA DEL HOLDING

En el gráfico 9 están resumidos los cambios que se producen entre un grupo de personas que el 2016 no contaba con un contrato de tarjeta del Holding, y durante el 2017 fue separado entre aquellos que abrieron su tarjeta y aquellos que no. Así, el grupo 1 es aquel grupo que abrió la tarjeta durante el 2017 y el grupo 0 es constituido por aquellas personas que no lo hicieron. Al igual que para el caso siguiente se separan los clientes en rangos de gasto para hacer comparables los resultados en términos de lift, usando la siguiente definición:

$$Lift_i = \frac{Gasto_{2018,i}}{Gasto_{2016,i}} \qquad o \quad Lift_i = \frac{Contribución_{2018,i}}{Contribución_{2016,i}}$$

Fórmula 1: Lift del cliente i

En todos los rangos y de forma transversal el grupo que abrió la tarjeta del Holding tuvo un mayor crecimiento de gasto entre los años 2016 y 2018. Otro resultado interesante es que, si bien en los rangos más bajos el grupo que no abrió tarjeta comercial tuvo un aumento en el gasto, esta variación se vuelve negativo en los rangos de gasto más

grandes. Esto no pasa para el grupo de personas que realizó el hito pues se observan lifts mayores a 1 en todos los rangos de gasto. Esto se repite consistentemente para la contribución, aquellos clientes que abrieron la tarjeta tuvieron un aumento en la contribución bruta. Se puede concluir que existen diferencias en el comportamiento de aquellas personas que realizan este hito y aquellas que no lo realizan.

Gráfico 9: Crecimientos en gasto y contribución promedio para grupos que abren tarjeta del holding (en miles)

2.4.4 COMPRA WEB POR PRIMERA VEZ EN CANAL WEB

En el gráfico 10 se puede observar las personas que durante el año 2015 y 2016 no contaban con compras web, pero si en físico. Aquí el grupo 1 fue el que realizó su primera compra web durante el año 2017 y el grupo 0 corresponde a aquellos que no. Se separó a las personas en rangos de gastos para que fueran comparables los porcentajes de aumento de gasto entre el año 2016 y 2018. Como se observa en todos los rangos de gasto durante el 2016, se produce que el grupo que compra por primera vez online tiene un lift mayor que los que no

Una de las características encontradas en el gráfico es que, a partir del rango de 600 mil pesos, todos los grupos de clientes tienden a disminuir su gasto durante el 2018, pero los que realizaron el hito tienen una disminución de gasto menor. Estos mismos resultados son homologables para el caso de la contribución, donde se muestra que aquellos grupos que disminuyeron su gasto lo hicieron menos si es que esos grupos realizaron su primera compra web durante el 2018. También se ve que los aumentos de contribución en

promedio de aquellas personas que realizaron el hito son mayores que las que no lo hicieron. Esto representa una oportunidad, pues se ve que al menos desde la perspectiva descriptiva existen diferencias en los comportamientos de compra de los grupos que realizan el hito de aquellos que no.

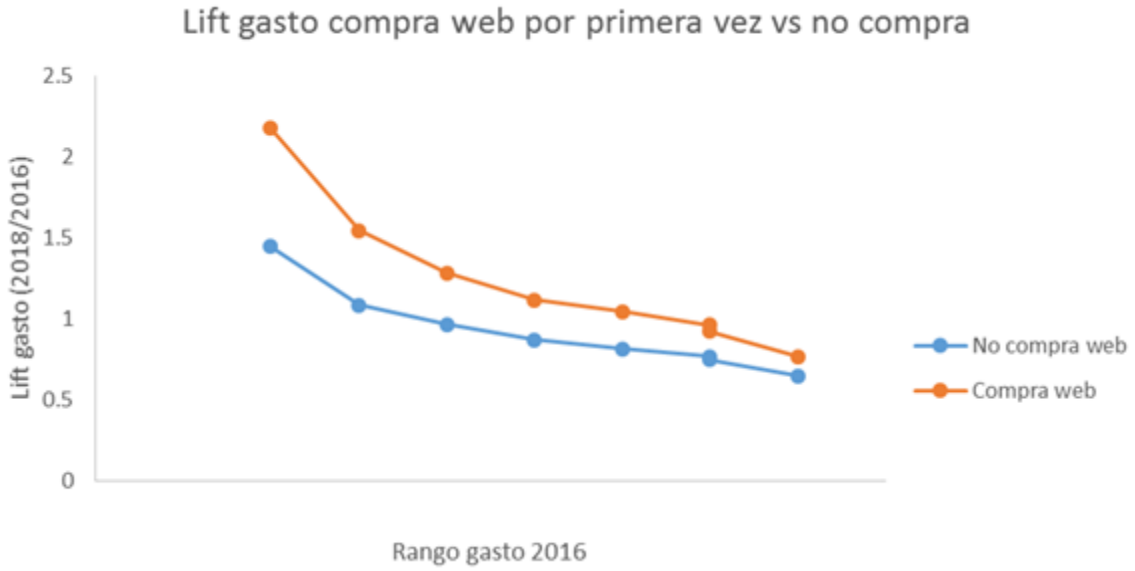


Gráfico 10: Crecimientos en gasto y contribución promedio para grupos que realizan primera
compra web (en miles)

2.4.5 Conjetura hipotética de valor agregado de la gestión

En base a estos 2 hitos, el valor de la oportunidad será la diferencia de gasto que se produciría si es que los clientes que no realizaron el hito tuvieran el mismo lift que aquellos que si lo hicieron, ponderado por un porcentaje

Para cada rango de gasto tomaremos el efecto de oportunidad como la diferencia del producto entre lift del grupo que realiza el hito con el gasto del 2016 del grupo que no realiza el hito y el mismo producto, pero considerando el lift real del grupo que no realizó el hito. Esto será un valor de la oportunidad promedio por cliente, por lo que después al multiplicarlo por la cantidad de personas que no realizaron el hito en ese rango de gasto se podría obtener un valor de la oportunidad

$$Valor_{oportunidad} = Gasto_{2016,0} * Lift_1 - Gasto_{2016,0} * Lift_0$$

Fórmula 2 : Valor de oportunidad (en gasto)

$$Gasto_{2016,0} = gasto\ durante\ el\ año\ 2016\ del\ grupo\ que\ no\ realiza\ hito$$

$$Lift_1 = Lift\ promedio\ de\ los\ clientes\ que\ realizaron\ el\ hito$$

$$Lift_0 = Lift\ promedio\ de\ los\ clientes\ que\ realizaron\ el\ hito$$

Como esta estimación descriptiva no toma en consideración el posible sesgo de autoselección por parte de los clientes, se asume de forma pesimista un 20% de éxito del aumento del lift. Al mismo tiempo que se hace esto es que se toma una cota de personas que son efectivamente gestionadas con los hitos y se toma como ejemplo el caso de que el 1% de las personas sean gestionadas efectivamente.

Al mismo tiempo que se hace esto, se toma una cota de personas que son efectivamente gestionados, esto con la finalidad de cuantificar la oportunidad de gestión

Hito	Gasto incremental	Contribucion incremental
Apertura tarjeta	$ 349,761	$ 214,211
Primera compra web	$ 517,799	$ 1,147,747
Total	$ 867,560	$ 1,361,968

Tabla 3: Ganancias del proyecto (dólares, a 680 CLP)

Por lo tanto, si se llegara a gestionar el 1% de los clientes de la muestra y considerando solamente un 20% del efecto descriptivo mostrado en la tabla 3 se podría lograr un aumento del gasto superior a 867 mil dólares y un aumento en la contribución bruta cercano a los 1.36 millones de dólares.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Aumentar el valor de la cartera de clientes de la empresa mediante el diseño de una herramienta de recomendación de hitos de gestión corporativo, con una perspectiva de valor corporativo.

3.2 OBJETIVOS ESPECÍFICOS

1) Definir y seleccionar de hitos a estudiar.

2) Estimar el efecto causal de la realización de cada hito en la rentabilidad de un cliente medido como la suma de las rentabilidades en el holding completo.

3) Estimar la probabilidad basal de cada cliente de realizar los hitos en ausencia de gestión ad hoc.

4) Generar recomendaciones comerciales de gestión a partir de la herramienta, para maximizar la utilidad de la cartera de clientes del holding

4. ALCANCE Y RESULTADOS ESPERADO

Los alcances son los siguientes:

- La rentabilidad de los clientes se medirá como la suma de las contribuciones brutas que generan las compras de estos, eso es, la resta entre el monto neto pagado por el cliente y el costo de adquisición del producto. No se incorporan otros conceptos como por ejemplo costos operacionales y administrativos.

- Solo se tomarán en consideración los costos e ingresos generados a través de los negocios de Formato 1, Formato 2 y Formato 3. Para los primeros 2 se considerarán tanto los canales físicos como web, en el último no se considerarán por contar con un número de transacciones muy bajo e imposibilidad de entrar a los valores de los costos de los productos. También, se dejan fuera los ingresos producidos por los servicios financieros, puesto que son de alto nivel de confidencialidad y no se cuenta con los privilegios suficientes para acceder a la información.

- Se utilizará la información transaccional histórica disponible en las bases de datos de la corporación, desde 2016 hasta marzo del 2019

- El modelo planteará recomendaciones de gestión, pero no incluirá experimentación pues para poder evaluar los resultados se tendría que esperar más de un año.

- Como entregable de esta memoria no se considera la puesta en marcha de la gestión corporativa mediante la herramienta diseñada.

- El trabajo de esta memoria no considera la recomendación de forma en la que se realiza las gestiones, por ejemplo, no se considera ver si es que la gestión tendrá que ser enviada al cliente por SMS o email o si descuento o algún otro tipo de enganche.

- Los hitos que formarán parte de las recomendaciones son: Apertura de tarjeta comercial, primera compra web en la corporación y cruce entre negocios en 2 variantes: desde Formato 1 (F1) a Formato 2 (F2) y Formato 2 a Formato 1

- Los clientes que contarán con recomendaciones serán personas naturales y no empresas.

5. MARCO TEORICO

En este marco teórico se muestran los principales modelos teóricos utilizados para resolver las problemáticas identificadas en la metodología.

5.1 ESTIMACIÓN DE EFECTOS CAUSALES:

Un efecto causal es el efecto generado por un tratamiento o una manipulación de la realidad sobre el valor de una variable de interés (Rajeev H. Dehejia and Sadek Wahba, 2002), que en el caso de esta memoria es la rentabilidad de los clientes para el holding.

La principal problemática de la estimación de los efectos causales es que para poder saber cuánto es el efecto de un tratamiento sobre una variable de interés en una persona se tiene que saber el valor de la variable de interés para persona la i en el tiempo t en presencia del hito y también el valor sin el hito. Por lo tanto, para la persona i en el tiempo t sobre la variable de interés Y su efecto causal sería:

$$\tau_i = Y_{i|hito} \; - \; Y_{i|no\ hito}$$

Fórmula 3: Efecto causal del hito

En general los efectos causales pueden tener varias métricas de medición una de estas es el ATT (Efecto promedio sobre los tratados).

Donde el ATT será:

$$\tau_{T=1} = E(Y_{i|hito} \; |T = 1) - \; E(Y_{i|no\ hito} \; |T = 1)$$

Fórmula 4: ATT

Donde T = 1 significa que el individuo fue en definitivo asignado al hito, como se mencionó anteriormente cada persona puede ser asignada o no al hito, pero no ambos estados a la vez, por lo que la identificación simultanea de los 2 componentes del efecto causal es imposible para cada cliente.

Para sortear esta problemática se puede estimar el valor esperado de la variable de interés en ausencia del hito para las personas designadas al hito. Este estimador es la esperanza del valor de interés para las personas que no recibieron el hito cuando la asignación de tratamiento fue realizada con asignación aleatoria.

Como este es un estudio observacional de datos históricos donde las personas son las que se autoseleccionan para la realización del hito, no se puede asegurar que existió asignación aleatoria y se necesitará aplicar otro método que resuelve este problema.

5.1.1 MATCHING

Para solucionar el problema de la inexistencia de asignación aleatoria es que se crea el matching. La metodología se basa en comparar sujetos que sean lo más parecidos posibles en alguna métrica de distancia, para después promediar estas diferencias y obtener una estimación del ATT antes presentado. Existe un amplio número de métricas de distancia que han sido utilizadas en la literatura, en este trabajo se utilizará la llamada Propensity Score Matching propuesto por Rubin y Rosenbaum en el año 1983 (Rosenbaum & Rubin, 1983).

Matching implica emparejar personas similares según características observables para hacerse cargo del problema del desbalanceo de variables (grupos de control y tratamiento distintos en sus distribuciones de variables descriptivas) para obtener una estimación insesgada del efecto del tratamiento (Rajeev et al. 2002). Cuando se cuenta con pocas variables observables de los clientes entonces el método de matching es directo, pues solo basta con mirar aquellos clientes que se parezcan en estas pocas variables. Pero cuando se tiene un gran número de variables observables de los clientes, entonces el método requiere de transformaciones para poder realizar el emparejamiento. En este contexto es donde más conviene usar el método de **Propensity Score Matching** (PSM) según los autores.

Por esta razón es que se decide utilizar el PSM en la memoria, pues la corporación cuenta con bastantes variables observables de los clientes. Por otro lado, el PSM fue planteado en 1983 (Rubin, 1983) donde se plantea que el balance de las covariables de los grupos de control y tratamiento es robustamente similar cuando se utiliza matching mediante propensity score, además de hacerse cargo de la complejidad de contar con muchas covariables (reduciendo la dimensionalidad a una sola).

En síntesis, el matching es una metodología en la que se excluyen del análisis a los clientes del grupo de control que no son comparables con alguno de los clientes tratados. De esta forma, a cada cliente tratado se le elige uno o más clientes controles que cumplen con los criterios que se explicarán más adelante según el método de Propensity Score Matching

5.1.2 PROPENSITY SCORE MATCHING

En el Propensity Score Matching (PSM) todas las covariantes de un individuo son reducidas a un escalar: la propensión a ser tratado. Según autores de la literatura de este método (Rajeev, et al 2002), se pueden utilizar un modelo probit o logit para esta labor. Para poder concluir que el match se realizó bien, se debe cumplir una condición importante: después de haber realizado el matching, los clientes control sobrevivientes deben tener la misma distribución agregada en sus variables centrales que los del grupo de tratamiento. Por lo que esto se debe testear y se deben utilizar medidas de desempeño que evidencien la similitud entre las distribuciones. En esta memoria se comparan las medias de ambos grupos antes y después del match para cumplir con esta tarea.

5.1.3 CONSIDERACIONES IMPORTANTES EN EL MATCHING

Existen consideraciones que no son parte del modelo en sí mismo, pero constituyen decisiones que se deben tomar en cuenta cuando se realiza el matching con cualquier métrica. A continuación, se resumirán las consideraciones expuestas por Rajeev et al.

- **Emparejar con reemplazo:** esto significa que, si es que un control que fue emparejado con un tratamiento puede ser usado nuevamente para ser emparejado con otro sujeto tratamiento. Lo anterior puede reducir la distancia de propensity score entre los sujetos tratados y no tratados, básicamente porque tendremos más posibles emparejamientos que en el caso sin reemplazo y cada unidad de tratamiento puede ser emparejada con su control más cercano. Esto es benéfico en reducción de sesgos, pero hacerlo sin reemplazo puede mejorar la precisión de los estimados.

- **Cuántas unidades de control emparejar a cada tratamiento:** seleccionar por ejemplo solamente una unidad de control por cada tratamiento implica que serán comparadas las unidades con el propensity score más cercano posible, lo que mejora el sesgo, pero empeora la precisión de la estimación.

- **Método de selección de unidades de control:** Los métodos más utilizados para emparejar son, los vecinos más cercanos (KNN), donde se elige un número específico de unidades más cercanas a los clientes activados. Otra variación es el método de radio de Caliper, donde se les exige a los clientes que se elijan para como emparejados que estén dentro de un radio máximo del cliente activado.

5.1.4 EFECTOS POR SEGMENTOS

Una vez matcheados los clientes, se puede estimar el efecto causal promedio de los hitos. Para poder obtener efectos heterogéneos se puede hacer segmentaciones de los clientes a través del propensity score. Esto se conoce como estratificación, donde los clientes son divididos en bloques según su propensión y en cada bloque se debe comprobar que las medias de los propensity score de ambos grupos es el mismo (Imbens & Rubin, 2009).

El método que se utilizará será el de clusterización, donde se armaran grupos basados en su cercanía en propensity score. Posteriormente a armar los segmentos se debe verificar el balance en cada uno de los segmentos, se proponen 3 métodos que se explicarán a continuación y se seleccionará el de menor SMD promedio entre los segmentos. Se propone que sean 5 segmentos tomando como criterio lo utilizado en anteriores memorias y la gestionabilidad.

5.1.4.1 K-means

K-means es un algoritmo de clustering cuyo objetivo es particionar una muestra de observaciones (clientes) en k distintos grupos, en el cual cada observación pertenece al cluster al cual su centro se encuentra más “cerca” (Francis Bach, lecture 3 2014)

Los K clusters se encuentran de tal forma que al realizar todas las asignaciones de cliente-cluster se haya minimizado la suma de todas las distancias entre cada observación y la media de su cluster en todas las variables utilizadas para la segmentación. En este caso la medida de distancia se hace por sobre una única variable, la propensión del cliente a activarse en el hito.

5.1.4.2 Quintiles

Este método es más simple que el caso del K-means porque de forma arbitraria se busca generar segmentos que conformen cada uno el 20% de la muestra total de clientes, posterior a haberlos ordenados de mayor a menor en su propensión a realizar los hitos. En la ilustración número 3 se observa un ejemplo de la segmentación mediante quintiles. En la ilustración, cada color representa un segmento distinto.

Ilustración 1: Segmentación de los clientes mediante quintiles

5.1.4.3 Cortes aleatorios

Este método corresponde a una generalización del caso de los quintiles, pues se construyen los clusters de forma aleatoria, donde se realizan cortes en distintos puntos desde el mínimo de propensión hasta el máximo, generando 5 segmentos. El set de cortes a utilizar corresponderá a aquel que entregue el mejor balance medido en SMD para las variables más importantes de los clientes. De todas formas, se pone como exigencia que cada segmenta tenga a lo menos un 10% de la muestra, con la finalidad de no terminar con segmentos muy pequeños que compliquen la propuesta de gestión de la memoria. La iteración se realiza la mayor cantidad de veces posibles, pero se propone una cota de 3000 iteraciones.

La selección final es aquella que de los 3 casos tenga el mínimo de SMD promedio entre los distintos segmentos. Se espera que el caso de cortes aleatorios sea el mejor, pues quintiles es un caso específico de cortes aleatorios y el algoritmo de K-means no guarda relación directa con las métricas de balance de las variables, además de ser un algoritmo hecho para segmentación en base a una mayor cantidad de variables.

5.2 ESTIMACIONES DE PROPENSIÓN

5.2.1 Regresión logística

Uno de los métodos utilizados para medir la propensión a realizar los hitos es el Logit, que busca encontrar una relación lineal que logre explicar las decisiones binarias de las personas en función de sus variables observables, calculando una probabilidad de que se realice una acción específica. En general los modelos probit se ajustan mediante la máxima verosimilitud (Lecture #11, Harvard University Fall 2016). La distribución

poblacional de probabilidades que se asume en el logit tiene la forma de la fórmula 5, donde los parámetros a estimar son los β's.

$$P(Y_i = 1) = \frac{e^{\beta_0 + X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

Fórmula 5: Logit

Y_i = 1 si el cliente i es tratado, 0 si no
β_1, β_0 = coeficientes a estimar
X_i = Vector de atributos del cliente i

En general la forma de estimar estos modelos es mediante la maximización de verosimilud (Lecture #11, Harvard University Fall 2016). Maximizar la verosimilud es optimizar los parámetros que predicen la decisión a través de la ecuación de forma que se maximice la probabilidad de ocurrencia de los datos observados.

5.2.2 Random Forest

El modelo de Random Forest corresponde a un modelo de ensamblaje, porque el resultado final de la predicción es el promedio de los resultados individuales de otros modelos. En este caso un Random Forest corresponde a un conjunto de árboles de decisión los cuáles votan cada uno para decidir la clase final (activada o no activada) que se predice. Un árbol de decisión es un modelo que va dividiendo la muestra de datos con criterios en base a las variables disponibles, donde cada división corresponde a un nodo. Se construye entrenándolo en un set de datos el cual en cada nodo hace un Split de los datos basado en la variable que mejor separe las clases de los clientes. Para tomar la decisión se utiliza una de las 2 posibles métricas: Gini o Entropy.

El lado aleatorio de un Random Forest se produce por el hecho de que para la construcción de cada árbol no se utiliza la muestra completa ni se consideran todas las variables para cada Split. Se toma una muestra aleatoria de los datos para construir el árbol y en cada nodo se toma un número definido (y menor al total) de variables de forma aleatoria en las cuales se evalúa el mejor split (Roxane Duroux, Erwan Scornet). Por consiguiente, el propensity score se obtiene como el promedio de las decisiones de los árboles de decisión.

5.2.3 Red neuronal

Las redes neuronales intentan replicar el funcionamiento de la red de propagación en las conexiones nerviosas de las neuronas humanas. En las cuales un mensaje se propaga

en forma de reacciones químicas a través de una red de neuronas hasta llegar a una respuesta final.

Así las redes neuronales artificiales cuentan con nodos (o neuronas) que se organizan en capas (layers), cada neurona de una capa se conecta con las neuronas de la capa siguiente. Existen 3 tipos de capas de neuronas la de entrada, las ocultas y la de salida donde se entrega el resultado del modelo. Así, cada observación de los datos de un cliente entra en una neurona input que se encuentra en la primera capa, cada input es pasado a todas las neuronas siguientes ponderada por un peso, por consiguiente, existe una ponderación para cada par de conexión.

$$a_m = \sum_{m=0}^{M} w_{km} x_k$$

Fórmula 6: Señal de llegada a neurona m

En la fórmula 6 se presenta la señal de entrada a la neurona m, donde w_{km} representa el peso que se le da a la señal particular de la neurona k de la capa anterior. Entonces la señal que llega a la neurona m es una combinación lineal de las señales de salida de todas las neuronas de la capa anterior a la que se encuentra la m. El mecanismo de aprendizaje de la red neuronal puede hacer que algunos pesos que conectan las neuronas sean 0.

Posteriormente esta neurona puede pasar su mensaje (la combinación lineal de los mensajes anteriores) dependiendo de una función de activación que se decide a priori de la calibración. En este caso se utilizará la función relu para las capas internas).

La función relu o rectificadora corresponde a $f(x) = (x, 0)$, esto implica que la neurona se activa y pasa su mensaje solo si este es positivo. El mensaje transmitido a la capa siguiente corresponde a la combinación lineal mostrada en la fórmula 6. De esta forma las señales de las neuronas se propagan desde la capa de input hasta la capa de output, en aquella capa solo hay una neurona y dadas las características del problema (modelamiento de clasificación) esta neurona tiene que tener una función ad hoc al problema que se quiere resolver: estimar los propensity score de los clientes. Por lo tanto, la función de activación en esta neurona toma la misma forma que en la de la regresión logística, correspondiendo a una función sigmoidea, haciendo que la combinación lineal final que llega a esta neurona se transforma en un output entre 0 y 1 (la propensión).

Este tipo de modelo aprende de los datos con los que se calibra con un método llamado backpropagation donde los parámetros a optimizar son los pesos w que unen los

mensajes de todas las neuronas. Se minimiza una función de costos en un número limitado de iteraciones.

5.3 PROPUESTA DE RECOMENDACIONES

Uno de los métodos de selección de clientes a activar en distintos tipos de hitos que se usa comúnmente en el retail es el de seleccionar aquellos clientes que son más propensos a realizar los hitos. Con este método se maximiza la efectividad de la herramienta de campaña, como puede ser un descuento o promoción. Esto sucede porque bajo el supuesto de que la propensión de un cliente de activarse en la actividad subyacente de la campaña está bien calculada, entonces se estará gestionando a los clientes que se sabe a priori es más probable que hagan lo que se gestiona.

No obstante, una gestión con esta perspectiva se queda corta en el análisis del valor. Pues, primeramente, no se sabe cuál es realmente el valor de corto, mediano o largo plazo de la gestión misma, sino que solamente se espera que el descuento o promoción utilizados, deriven en compras directas ligadas a la misma gestión. Así se deja de lado lo que se genera en el comportamiento del cliente a futuro. Por ejemplo, si es que gestionamos a los clientes para comprar por primera vez en el canal web y medimos la efectividad de esta gestión solamente basado en las compras que se produjeron inmediatamente por la gestión, entonces no se sabrá los posibles efectos futuros que podría generar en el cliente el hecho de que ya haya comprado en el canal web por primera vez.
Por consiguiente, se toman en consideración los puntos anteriores y se estimará que la rentabilidad esperada de la realización de un hito para cada cliente será: la multiplicación entre la propensión y el efecto causal del segmento al cual pertenece el cliente. La fórmula 7 muestra la rentabilidad esperada de que el cliente i realice el hito j

$$Rentabilidad\ esperada_{i,j} = \left(P_i\left(hito_j\right)\right) * Efecto\ causal\ _{i,j}$$

Fórmula 7: Formula de rentabilidad esperada de la gestión de los hitos

$$P_i\left(hito_j\right) : Probabilidad\ de\ que\ cliente\ i\ realice\ hito\ j$$
$$Efecto\ causal\ _{i,j} : cliente\ i\ realice\ hito\ j, calculado\ según\ el\ segmento\ al\ cual\ pertenece$$

Así, como muestra la fórmula 7, la rentabilidad esperada de un cliente es creciente en la probabilidad de realización del hito, pero también puede pasar que, al pertenecer a un segmento de mayor propensión, entonces los efectos causales vayan cambiando.

Finalmente, para realizar las recomendaciones, se estimará la base potencial de clientes a gestionar (o sea los clientes que aún no se activan) y se calcula la rentabilidad esperada

por segmento de cada hito, con la finalidad de discriminar y priorizar a segmentos gestionar y decidir con que montos de inversión máximos en descuento o cupones.

5.4 METRICAS DE DESEMPEÑO

5.4.1 MÉTRICAS DE LOS MODELOS E PROPENSIÓN

Las métricas de desempeño describen que tan buenos clasificadores son los modelos utilizados en la memoria, con respecto a la capacidad de diferenciar entre los clientes que se activan y los que no se activan. Las métricas son las siguientes:

5.4.1.1 MATRIZ DE CONFUSIÓN

La matriz de confusión resume los números de los clientes clasificados poniendo los resultados agregados en 4 cuadrantes: Clientes activados clasificados correctamente (True Positive), Clientes activados clasificados erróneamente (False Negative), Clientes no activados clasificados correctamente (True Negative) y los clientes no activados clasificados erróneamente (False Positive). La tabla se resume en la ilustración 2.

		Clase real	
		Activado	No activado
Clase predicha	Activado	True Positive	False positive
	No activado	False negative	True negative

Ilustración 2: Matriz de confusión

Sin embargo, para la construcción de estos cuadrantes se debe tomar una decisión sobre cuál será el mínimo de probabilidad para considerar a un cliente como activado. Si bien este corte puede parecer arbitrario, estas métricas son consideradas a la hora de comparar distintos modelos en un mismo caso de negocio, comparando con la misma arbitrariedad a todos los modelos. Esto implica que las métricas que salen de esta matriz de confusión no servirían para hablar acerca del desempeño de un modelo de clasificación de forma aislada y solo sería correcto utilizarlas cuando se comparan modelos en un mismo contexto.

5.4.1.2 ACCURACY

El accuracy es el cociente entre los clientes que fueron catalogados correctamente como activados o no activados y la totalidad de los clientes. Esta métrica es un primer acercamiento al desempeño de los clasificadores pues habla en forma promedio de qué porcentaje de la muestra fue clasificada correctamente. La fórmula 8 muestra la expresión.

$$Accuracy = \frac{True\ posive + True\ negative}{Toda\ la\ muestra}$$

Fórmula 8: Accuracy

5.4.1.3 RECALL

El recall es el cociente entre los clientes activados predichos correctamente y todos los clientes activados. Es una especie de accuracy considerando solamente a los clientes que estaban realmente activados, por lo tanto, nos dice que tan capaz es el modelo de detectar a los activados. La fórmula 9 muestra la expresión

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$

Fórmula 9: Recall

5.4.1.4 SPECIFICITY

El specificty es el cociente entre los clientes no activados predichos correctamente y la totalidad de los clientes no activados. Este caso es el mismo que el recall pero para los clientes no activados, habla acerca de que tan capaz es un modelo de predecir a los no activados. La fórmula 10 muestra la expresión

$$Specificity = \frac{True\ Negative}{True\ negative + False\ positive}$$

Fórmula 10: Specificity

5.4.1.5 CURVAS

Existen otros tipos de métricas de clasificadores que no dependen necesariamente de donde se deja el límite por el cual se considera a un cliente como activado. Este es el caso de las curvas CAP y ROC-AUC. Estas curvas hablan directamente sobre la capacidad de los modelos de separar la muestra de clientes en aquellos que se activan

de los que no se activan. Por otro lado, al no depender del límite en el cual se consideran activados y no activados, estas métricas funcionan de mejor forma al considerarlas como métricas absolutas desempeño y no como métricas relativas.

5.4.1.6 CUMULATIVE ACCURACY PROFILE – CURVA CAP

La curva CAP ordena a los clientes desde mayor probabilidad predicha a menor probabilidad predicha y se hace el análisis siguiente, qué porcentaje de los activados se logra capturar a medida que se designa como activados todos los clientes, partiendo desde los más probables a los menos probables. El modelo base es aquel que designa con igual probabilidad a activados y no activados. Un ejemplo de esto es lo que se muestra en la siguiente figura:

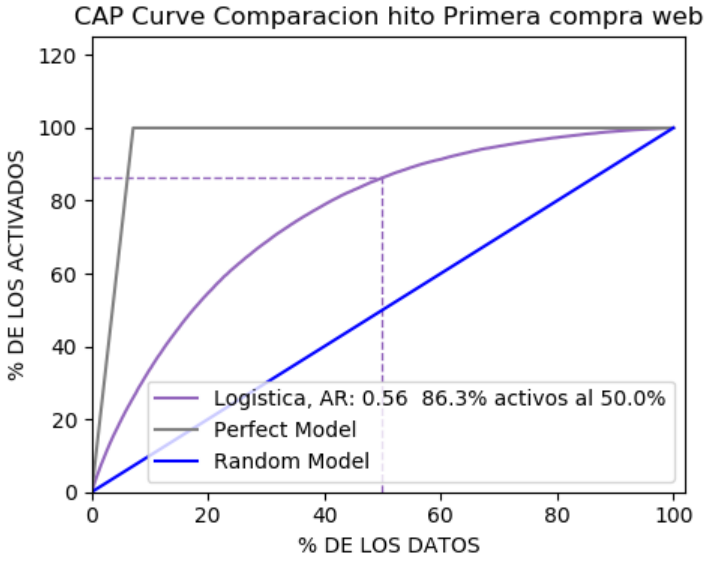


Gráfico 11: Curva CAP ejemplo hito primera compra web

En el gráfico 11 se ve el ejemplo de la curva CAP para el hito primera compra web utilizando el modelo de regresión logística. En el gráfico se aprecia una muestra de clientes donde aproximadamente el 10% de ella son de la clase activado y el 90% pertenecen a clase de no activados. La línea gris representa el modelo perfecto, aquel modelo que separa perfectamente a los clientes activados de los no activados mediante el porcentaje de probabilidad, pues para poder capturar a todos los activados solo basta con contactar al 10% de la muestra total. El caso del modelo random se debe designar como activados a toda la muestra para poder predecir correctamente el 100% de ellos. Así el mejor modelo será aquel que logre identificar a un mayor porcentaje del grupo de activados contactando a la menor cantidad de personas.

Por otro lado, la línea púrpura representa el modelo a evaluar, se puede observar que un 86.3% de los activados son logrados identificar al llegar a la mitad de los clientes. Junto con esto también se observa la métrica AR, que representa la proporción del área entre la curva perfecta y el modelo aleatorio que logra cubrir el modelo utilizado. El análisis dice que el modelo con un AR mayor será el mejor discriminador entre clases de clientes.

5.4.1.7 RECEIVER OPERATING CHARACTERISTIS CURVA ROC - AUC

La curva ROC-AUC (Gráfico 12) es un análisis que permite entender la capacidad de separación de clases de un modelo. En esta curva se grafica el True Positive Rate (recall) y el False Positive Rate (1 – Specificity). El análisis corresponde a mover entre 0 y 1 el límite mínimo de probabilidad para clasificar un cliente como activado y se observa cómo varían las métricas en cada caso. A continuación, se muestra un ejemplo.

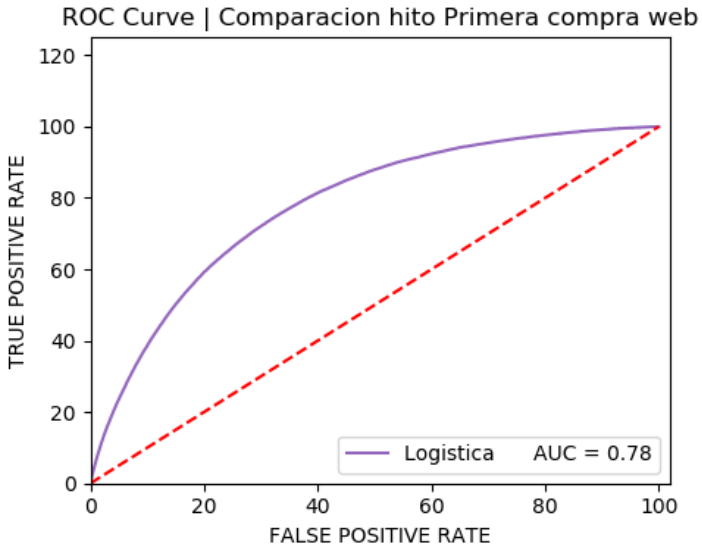


Gráfico 12: Curva ROC-AUC ejemplo hito primera compra web

Este gráfico se interpreta pensando que pasaría con el Recall si es que se mueve el límite de activado entre 0 y 1. Así, en el extremo izquierdo tenemos el caso donde el límite es 1, o sea que los clientes con probabilidad mayor a 1 serán considerados como activados. El Recall del caso anterior es 0 y el False Positive Rate también, esto pues ningún cliente es considerado como activado. El otro caso es el extremo derecho donde el límite es 0, por lo tanto, todos serán considerados como activados y el Recall se vuelve 1, pero también se vuelve 1 la tasa de no activados clasificados erróneamente puesto que ningún cliente será clasificado en esa clase. La línea punteada representa el conjunto de puntos donde la tasa de verdaderos positivos y falsos positivos es la misma.

En resumen, el análisis ROC-AUC nos permite responder la pregunta de qué tan bien el modelo es capaz de identificar a los activados en medida que sacrificamos predecir erróneamente a los no activados. Por lo tanto, el modelo será mejor en medida que pueda identificar a los activados sin perder la capacidad de predecir a los no activados, esto se mide como el área bajo la curva de la línea purpura en este caso, que vendría siendo 0.78.

5.4.2 MÉTRICAS DEL MATCHING

5.4.2.1 BALANCE DE VARIABLES

El balance se refiere a que ambos grupos, los activados y los no activados, tengan las mismas distribuciones de sus variables explicativas más importantes. Es posible hacer inferencias cuales cuando ambos grupos están balanceados en sus atributos, lo que se puede lograr utilizando el Propensity Score como herramienta de balance (Peter C. Austin, 2009)

La métrica más comúnmente utilizada por los investigadores para evaluar el balance entre las variables es la diferencia de medias estandarizadas (SMD) (Elizabeth A. Stuart, et al., 2013). La diferencia de medias estandarizadas compara las medías de ambos grupos, en unidades de desviación estándar conjunta de ambos grupos. Así, para medir el balance de las variables entre ambos grupos se utilizará esta métrica sobre un conjunto de variables que son consideradas importantes. Para poder seguir adelante con el proceso de estimación de efectos causales es que cada una de estas diferencias debe estar bajo 0.1. Una diferencia de este nivel es considerada indistinguible y suficiente para un considerar un match como satisfactorio (Normand et al., 2001).

A diferencia de los test de hipótesis que se utilizan para comparar las medias con algún nivel de significancia, la SMD no se ve influenciada por el tamaño de las muestras de los grupos a comparar. Esto resulta relevante, pues para estimar los efectos causales en clusters, la muestra es separada en grupos más pequeños. También es importante considerar que ésta métrica de balance hace comparables variables en distintas escalas, como podrían ser gasto que puede llegar hasta cifras de millones y edad que no supera los 100 años. Las fórmulas 11 y 12 muestran la expresión matemática de la métrica

SMD en variables continuas:

$$d = \frac{\mu_{activado} - \mu_{no\ activado}}{\sqrt{\frac{s^2_{activado} + s^2_{no\ activado}}{2}}}$$

Fórmula 11: SMD para variables continuas

μ: Media de la variable en cada grupo
s: Desviación estándar de cada grupo

SMD en variables binarias:

$$d = \frac{P_{activado} - P_{no\ activado}}{\sqrt{\frac{P_{activado}(1 - P_{activado}) - P_{no\ activado}(1 - P_{no\ activado})}{2}}}$$

Fórmula 12: SMD para variables binarias

P: Proporción del grupo en que la variable es 1

Este cálculo se hará solamente para las variables más importantes en la explicación de la propensión de un cliente a realizar los hitos. Esto considerando la posible existencia de variables que no sea relevantes en la explicación de los hitos y que podrían sesgar nuestro resultado. En caso de que alguna variable importante quede por sobre el 0.1 en valor absoluto entonces se debe volver a la modelación del modelo de propensión y empezar a considerar agregar interacciones entre variables o cambiar sus formas funcionales, probar los valores al cuadrado o en raíz cuadrada.

5.5 HERRAMIENTAS DE MODELACIÓN

5.5.1 ESCALAR VARIABLES

Los datos de una muestra pueden ser reescalados mediante alguna fórmula para obtener beneficios con respecto a la facilidad con la que los modelos resuelven sus propias funciones de costos. Algunos beneficios de estas transformaciones corresponden a:

- Disminuir el número de iteraciones necesarias para que los modelos converjan a una solución óptima según su algoritmo de aprendizaje y función de costos.

- Interpretabilidad de parámetros en aquellos modelos que los tengan. Al escalar las variables los parámetros de cada variable se vuelven comparables entre sí porque se moverán dentro de los mismos órdenes de magnitud (Ejemplo: edad versus gasto)

El método mediante el cual se reescalan las variables en esta memoria es la estandarización. Con aquel método los datos de las variables de un cliente son

reemplazados por su valor menos la media de la muestra dividido en la desviación estándar, según la siguiente formula (según muestra la fórmula 13). Como resultado la variable estará centrada en 0 con una desviación estándar de 1.

$$X_{escalado\ de\ i} = \frac{x_{\ i} - \underline{x}}{\sigma_x}$$

Fórmula 13: Normalización de variable x

$$\underline{x}: Promedio\ muestral\ de\ X$$
$$x_{\ i}: Valor\ de\ la\ varible\ x\ para\ el\ cliente\ i$$
$$\sigma_x: Desviación\ estándar\ muestral\ de\ X$$

5.5.2 CROSSVALIDATION

Mediante el método de k-fold Crossvalidation la muestra que se utilizará como input en un modelo es dividida en un número K de segmentos de forma aleatoria (en casos de series de tiempo con dependencia Inter temporal deben ser segmentos de datos ordenados en tiempo). Posteriormente el modelo utilizado es entrenado en k-1 de los segmentos del CrossValidation y su poder de predicción es puesto a prueba en el segmento que no se utilizó para la calibración. Este proceso se hace k veces, hasta tener los resultados de las k iteraciones. Posterior a esto se elige el mejor modelo del CrossValidation y se calibra en todos los datos, finalmente este modelo se usa para predecir en un conjunto de datos que no fueron utilizados para el CrossValidation y se confirma si es que no existe overfitting de los modelos.

El overfitting no es deseable en contextos predictivos porque es un síntoma de pérdida de capacidad de generalización de los modelos y sesgo de los resultados. En ese caso los modelos no se comportan de buena manera en datos que el modelo no utilizó para aprender.

5.5.3 GRIDSEARCH

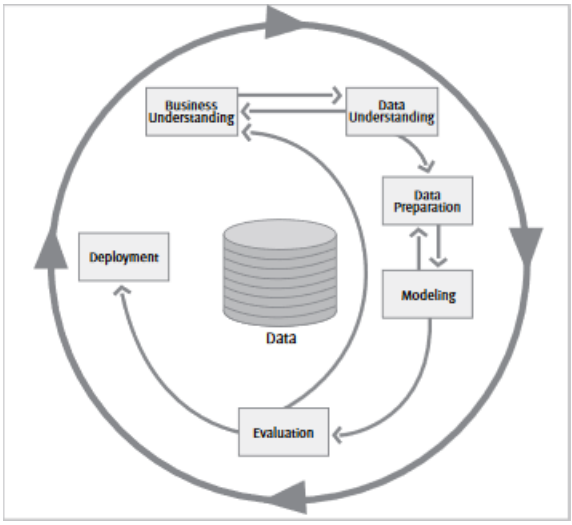
El GridSearch o búsqueda en grilla es un método por el cuál un modelo prueba su desempeño con distintos hiperparámetros para encontrar la mejor combinación de ellos en las métricas de desempeño.

Los hiperparámetros son aquellos parámetros que no se encuentran mediante la calibración de un modelo sobre un set de datos, sino que se deciden a priori de la calibración. Como se deciden a priori la única forma de encontrar los mejores hiperparámetros es iterando los modelos sobre los datos usando distintos valores para estos parámetros como input. Así la grilla corresponde a las combinaciones de parámetros que se le entregan como casos posibles a probar al método de GridSearch. Como ejemplo, el modelo Random Forest puede ser iterado con distintos valores de: Cantidad de árboles en el bosque, máxima profundidad de cada árbol, cantidad máxima de variables que se toman de forma aleatoria a la hora de evaluar cada Split de datos, entre otros.

6. METODOLOOGÍA

Existen dos marcos metodológicos que son los más utilizados en los proyectos de ciencia de datos, el primero es KDD y el segundo es CRISP-DM. A diferencia del KDD, el CRISP-DM cuenta con una fase inicial que corresponde al entendimiento del negocio (CRISP-DM 1.0, SPSS 2000). Justamente por la existencia de esta primera fase de entendimiento del negocio es que se elige utilizar este marco metodológico, pues se entiende que al intentar generar una nueva propuesta que propone pensar un nuevo tipo de gestión que piense a los clientes de forma corporativa, es que se necesita entender en cierto nivel el negocio por las posibles implicancias que tenga en la propuesta final de esta memoria

Se hará una explicación de la metodología mediante los mismos pasos propuestos por el esquema de CRISP-DM usando de referencia la guía de uso gecha por SPSS en los años 2000, pero aterrizándolo a la realidad de este proyecto.



Esquema 1: Esquema del modelo CRISP-DM (fuente: CRISP-DM 1.0, SPS 2000)

6.1 ENTENDIMIENTO DEL NEGOCIO

Esta fase se dedica al entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva de negocios, para transformar este requerimiento en un problema de minería de datos y plan preliminar diseñado para conseguir estos objetivos.

En esta fase se debe investigar acerca de cuáles son las fuentes de ingresos en cada negocio, entender las características fundamentales de los negocios y si es que es relevante, desde una perspectiva corporativa, la gestión de los clientes, de esta fase nacen las definiciones fundamentales de lo que consiste esta memoria.

6.2 ENTENDIMIENTO DE LOS DATOS

Aquí se comienza con un primer inicio de recolección de data para proceder con actividades que hagan que el autor se familiarice con la información disponible, identificar problemas en la calidad de los datos, descubrir los primeros descubrimientos y encontrar o detectar submuestras de datos interesantes para formular hipótesis a cerca de información escondida.

Esta fase es relevante pues puede llegar a reevaluar los alcances de la memoria, resulta esencial en este caso, entender cuáles serán las fuentes de información acerca de las rentabilidades de los negocios, los pasos principales son:

- Entender si es que se cuenta con información de los costos asociados a las compras de los clientes.
- Cuáles son los costos asociados a las ventas son identificados en las bases transaccionales.
- Qué tipo de información cuentan las bases transaccionales.
- Ver si es que se puede diferenciar ventas de devoluciones o cambios.

Por otro lado, resulta importante saber si es que se tiene acceso a información de los clientes para la utilización de los modelos predictivos:

- Identificar distintas fuentes de datos.
- Revisar la existencia de sesgos en la tenencia de datos de los clientes, por ejemplo, se podría tener mejor calidad de datos de los clientes que tengan tarjeta comercial.

Para esta fase se tiene recopilar información de las fuentes a las cuales se tiene acceso, una vez hecho esto se tiene que ver la calidad de los datos mediante un de análisis descriptivo inicial, con las siguientes finalidades:

- Ver si es que hay variables que están bien pobladas donde un gran porcentaje de los clientes cuenta con ellas.

- Ver la presencia de errores o valores atípicos en los datos que podrían impactar la confiabilidad de los resultados.

Por otro lado, si bien los hitos son seleccionados mediante criterio experto por parte de miembros de la corporación, se debe realizar un análisis de la capacidad de identificación de estos hitos en los datos.

- Para el caso de apertura de tarjeta comercial, se debe observar si es que se cuenta con la información de la apertura de la tarjeta y el estado de la misma.

- Para el caso de primera compra web se debe investigar si es que está clasificado el canal por el cual se producen estas compras en las bases de datos transaccionales.

- Para todos los hitos se debe observar si es que se cuenta con número razonable de casos anuales, entendiendo que valga la pena la gestión de los mismos además de que asegure un buen desempeño de los modelos a utilizar.

6.3 PREPARACIÓN DE LOS DATOS

En esta fase se consideran todas las actividades para construir el set de datos final utilizado para los modelos. En esta fase se consideran todas las actividades del tipo selección de tablas, registros y atributos, al igual que transformación y limpieza de datos.

Esta fase decide cuáles serán las fuentes de datos a utilizar. Cada negocio cuenta con una base transaccional que registra data relevante para cada transacción y se debe decidir cuales tipo de transacciones se considerarán. Por ejemplo, se debe definir si se utilizarán ventas con facturas o solamente ventas con boleta, como también si se contabilizarán las devoluciones de los productos

También se debe identificar valores atípicos y en base al análisis definir reglas de exclusión de información. Por otro lado, se tiene que analizar la posibilidad y necesidad de aplicar algún tipo de transformación a los datos utilizados para calibrar los modelos, ya sea normalizar o estandarizar, esto porque existen modelos que tienen un mejor desempeño con datos transformados.

Otra decisión que se debe tomar es cómo tratar la información de los clientes no identificada (los valores nulos), donde las opciones son:

- Reemplazar con la moda

- Reemplazar con la media
- aproximación lineal en caso de ser atributos continuos

Finalmente, también está la posibilidad de no utilizar el atributo si es que se cree que modificar los valores de los clientes que tengan nulo sería perjudicial para el poder de modelación del proyecto.

6.4 MODELAMIENTO

En esta sección se seleccionan los modelos y se aplican al problema del negocio. A continuación, se muestran los principales componentes de la memoria que resuelven los objetivos específicos.

6.4.1 Rentabilidad esperada de la gestión del hito

Este será el output final de la memoria y tomará en consideración el efecto causal estimado para cada hito y las estimaciones de propensión de que un cliente realice el hito. Así la rentabilidad esperada será la multiplicación del efecto causal y la probabilidad de una persona de que realice el hito, como se muestra en la fórmula 14.

$$Rentabilidad\ esperada_{ij} = P(x_i) * Y(Hito_j|s_i)$$

Fórmula 14: Rentabilidad del hito j sobre el cliente i

$$Y(s_i) = Efecto\ causal\ del\ cliente\ i\ perteneciente\ al\ cliente\ i\ perteneciente\ al\ segmento\ i$$

$$P(x_i) = Probabilidad\ de\ que\ el\ cliente\ i\ realice\ el\ Hito\ j\ según\ sus\ características$$

6.4.2 Estimación de efectos causales

La relevancia de aplicar este tipo de modelamiento recae en el hecho de que no se quiere hacer un análisis descriptivo de los clientes que hacen un hito, sino que se quiere saber cómo es que la realización del hito afectó los patrones de consumo de los clientes. Esto con la finalidad de aplicar y gestionar en base a estos resultados a los clientes que aún no han realizado los hitos estudiados. Existen distintos métodos de estimación de efectos causales, el seleccionado se nombra en la sección de marco conceptual.

Así la recomendación de gestión se basará en la estimación del efecto causal que se espera que la gestión del hito tenga por sobre los clientes gestionados.

Otro punto que considerar es la estimación de efectos causales heterogéneos entre los clientes, lo que se realizará mediante segmentación de los resultados causales obtenidos

en la última parte de la metodología utilizada. Por lo tanto, se segmentará a los clientes de la firma para encontrar efectos causales por segmentos.

6.4.3 Modelo de propensión

La importancia de la estimación de la propensión de realizar un hito recae en que los efectos causales estimados en la modelación anterior deben ser ponderados por la probabilidad de que una persona efectivamente realice el hito. Lo anterior se realiza porque una gestión podría hipotéticamente tener un gran valor estimado de aumento de rentabilidad, pero a la vez tener una muy baja probabilidad de realización del hito por lo tanto el efecto esperado no sería tan alto.

Esta propensión debe ser calculada al momento de querer obtener las recomendaciones, pues esta probabilidad puede variar con el tiempo dependiendo en el comportamiento que tenga un consumidor. Ya que, parte de la propensión a realizar un hito se debe a variables sociodemográficas que en general no tienen gran variabilidad en el tiempo, pero otra parte será a causa de la historia transaccional del cliente. Y estas últimas pueden ir variando constantemente, dependiendo de la especificación de las variables.

6.4.4 Ventana de estimación de los efectos

Las ventanas temporales para medir los efectos causales se muestran ilustrados en la ilustración 3.

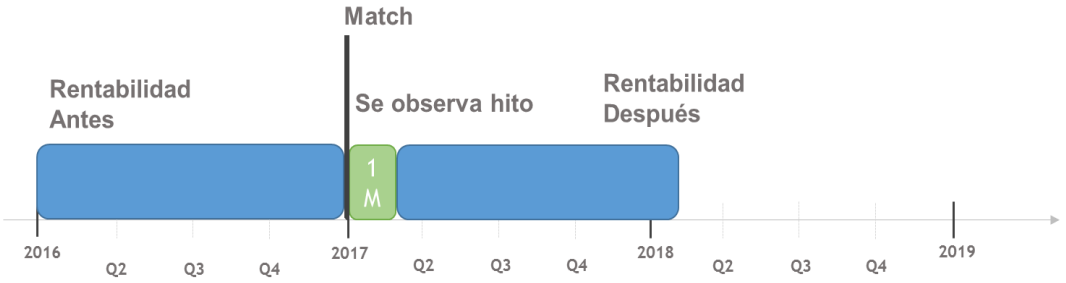


Ilustración 3: Ventanas temporales relevantes, ejemplo

En el esquema se puede ver que la ventana temporal de medición del efecto será de 12 meses, tanto antes como después de observar el hito. Estas ventanas temporales corresponden a lo que en el esquema se señala como “rentabilidad antes” y “rentabilidad después”

La otra ventana temporal es la de la observación de la realización del hito, esta corresponde al periodo en el que se observa si es que los clientes realizan el hito. De forma intuitiva una ventana de observación del hito de mayor duración implicaría necesariamente un número más grande de gente que realiza el hito como input de los modelos

6.4.5 Ventana de estimación de las propensiones finales

La ventana de estimación de las propensiones a utilizar para las recomendaciones de gestión será la de la ilustración número 4. Se puede observar que, con las variables relevantes de 1 año anterior a la observación del hito, se calibrarán modelos para predecir las propensiones actuales. Los datos observados en este caso serán de 1 meses en los que los clientes pueden realizar los hitos.

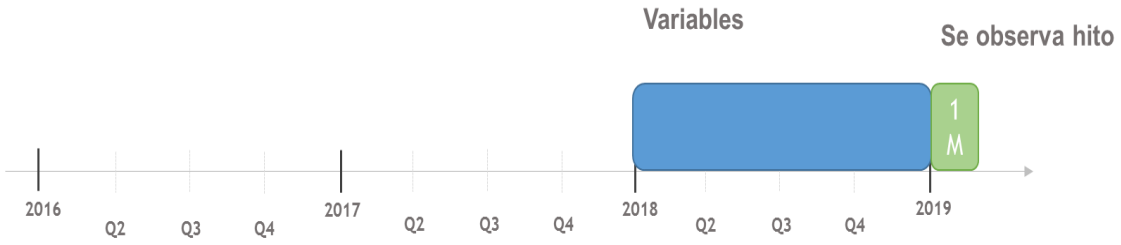


Ilustración 4: Ventanas temporales relevantes estimaciones propensiones finales

La ventana de observación de los hitos con los cuales se calibrarán los modelos finales a utilizar en la última etapa de las recomendaciones será también de 1 mes, pues debemos simular el mismo proceso con el cual se calcularon los efectos causales. Si bien existe la posibilidad de que al utilizar una ventana más amplia de tiempo se obtengan más clientes activados de los cuales aprender, también se pierde la consistencia del mismo proceso metodológico, pues si los efectos causales son estimados con una ventana de solo 1 mes, entonces se arriesga perder generalidad al cambiar la ventana posterior. Por lo que, para ser consistentes en la metodología, también se utilizará un mes para el modelo calibrado en el 2019.

Por consiguiente, la ventana de observación de los hitos para el modelo final de propensión será enero del 2019, con los cuales se recalibran los modelos de propensión de cada hito. Posteriormente con las variables calculadas entre 01 de julio del 2018 y 30 de junio del 2019 se estimarán las probabilidades de que los clientes se activen durante julio del 2019 y se analizará estos resultados en perspectivas de clientes potenciales.

6.5 EVALUACIÓN

Con los modelos ya creados y antes de pasar a la fase final se debe hacer una evaluación minuciosa del desempeño de la modelación y revisar cada uno de los pasos utilizados para crear la solución, para poder tener certeza de que la modelación resuelve el problema de negocio que motivó el proyecto. Un objetivo clave de esta fase es determinar si existe alguna importante consideración de negocio que no haya sido resuelta. Así en esta parte principalmente se realizarán los siguientes pasos:

- Revisar métricas de los modelos utilizados
- Revisar robustez de los resultados

Las métricas utilizadas para medir la efectividad de la modelación consistirán principalmente en ver el poder predictivo de los modelos de propensión y también evaluar si la medición de los efectos causales tiene suficiente heterogeneidad entre los clientes, pues la idea es tener ofertas personalizadas según segmentos. Las métricas serán explicadas en la sección del marco teórico.

6.6 DESPLIEGUE

El despliegue de la herramienta construida en esta memoria no será considerado parte del proyecto.

7. DESARROLLO DE LA METODOLOGÍA

7.1 DATOS UTILIZADOS

En esta parte del desarrollo de la metodología se habla de lo relevante con respecto a las distintas variables que se utilizaron para la estimación de los efectos causales de los 4 hitos que se trabajaron en la memoria. Se habla en forma general de los 4 hitos y en detalle sobre el hito de primera compra web.

7.1.1 RESUMEN DE DATOS

Las ventanas temporales utilizadas para la modelación del problema de estimación de efectos causales fueron las siguientes:

- Ventana anterior al hito: Desde enero 2016 hasta diciembre 2016
- Ventana observación del hito: enero 2017
- Ventana posterior al hito: Desde febrero 2017 hasta enero 2018

De la ventana anterior al hito se obtienen las variables que se utilizan para la estimación de las propensiones de los clientes, además de las variables de evaluación en el periodo anterior del hito. Las variables de evaluación son aquellas variables sobre las cuales se estima el efecto causal, en el caso de esta memoria están son: gasto, contribución y transacciones. Estas variables son calculadas sobre los tres negocios del alcance de la memoria.

En general en las variables utilizadas para los modelos de propensión tendremos gasto y transacciones en cada negocio y en distintas ventanas de tiempo. Las ventanas pueden ser 1 mes, 3 meses, 6 meses y 12 meses, separando canal físico y digital. Otras variables que se utilizan son las del tipo sociodemográfico, como edad, sexo, y región. Además, también existen variables relacionadas a la tenencia de productos financieros, como lo son: tenencia de tarjeta, tipo y cupo.

Por otro lado, existe un set de variables que son limitadas por los hitos a estudiar. Por ejemplo, para el caso del hito de apertura de tarjeta comercial, se tiene que las variables relacionadas a tenencia de tarjeta no existen pues se limita a la base de clientes aquellos que no han realizado el hito todavía. Caso similar al anterior sucede con la primera compra web, pues todas las variables que correspondan a transacciones en los distintos canales digitales se encuentran vacías por construcción del caso de estudio.

Así, en la ventana de observación del hito se registra si es que los clientes monitoreados realizan el hito. Primeramente, se cuenta con un set de clientes que justo en el tiempo anterior a enero 2017 no han realizado el hito en cuestión, el cual se divide en 2 durante la ventana de observación: aquí se define como 1 (activado) al cliente que realiza el hito y 0 (no activado) el que no lo realiza. Es importante mencionar que si un cliente no se activa en la ventana de observación, se puede activar con posterioridad y aun así se considerará como "0"

Durante la ventana posterior al hito se miden las variables de evaluación, con la finalidad de medir los cambios. Esto consiste en medir para cada cliente su crecimiento absoluto (resta simple) de las variables de evaluación. El estimador final del efecto causal será el promedio de las diferencias de crecimiento entre el activado y el no activado matcheado a su activado correspondiente.

De esta forma se tendrán set de clientes distintos para cada hito, un cliente que no haya realizado el hito durante la ventana anterior puede haber realizado otro hito y esto no se considerará como condición para sacarlo de la otra base. Cada base consiste en los clientes que cumplan las condiciones definidas a continuación:

- Hito primera compra web: No haber comprado en el canal web hasta un año antes a la ventana de observación del hito

- Hito apertura de tarjeta: No tener tarjeta comercial activa (que no esté suspendida ni castigada). en el mes anterior a la observación del hito, excluyendo a los clientes que hayan sacado una tarjeta adicional en el periodo de observación e incluyendo a los clientes que puedan haber cerrado su tarjeta anteriormente.

- Hito de cruce desde F1 a F2: Clientes que tengan al menos una transacción (en canal web o físico) en el negocio de Formato 1 y ninguna en el negocio de Formato 2.

- Hito de cruce desde F2 a F1: Clientes que tengan al menos una transacción (en canal web o físico) en el negocio Formato 2 y ninguna en el Formato 1.

Se toma como filtro inicial que los clientes que se utilicen en el desarrollo del trabajo tengan al menos una compra en cualquier negocio del holding. A continuación, se muestran las cantidades de clientes que cumplen esta condición en cada hito y separados en activados y no activados:

Hito	Porcentaje activado
Compra web	1.0%
Tarjeta	0.3%
Cruce de F2 a F1	3.4%
Cruce de F1 a F2	3.3%

Tabla 4: Clientes que se activan vs clientes que no se activan en cada hito

Como se observa en la tabla 4, se puede apreciar que existe una baja cantidad de clientes que realizan el hito en la ventana de observación, donde el caso más extremo es el de la apertura de tarjeta, donde solo un 0.3% de los clientes potenciales a realizar el hito lo realizan. Por otro lado, el cruce de negocios se realiza con mayor frecuencia que la apertura de tarjeta y la primera compra web.

Como criterio general se decide que las bases a utilizar en el desarrollo del trabajo sean constituidas por todos los clientes activados durante el periodo de observación y una muestra de los no activados que sea 10 veces la cantidad de los activados. Las razones para hacerlo de esta forma son principalmente por el intensivo requerimiento de recursos por parte de los modelos a utilizar y además de contar con varios candidatos posibles para el matching.

7.1.2 VARIABLES

Como se menciona en el apartado anterior, hay variables que son limitadas por la misma construcción del problema a estudiar, como, por ejemplo: las variables de tarjeta no tienen datos para aquellos clientes que no han realizado aun el hito de apertura de tarjeta. Tomando en consideración lo anterior a continuación se muestra la totalidad de variables a utilizar en los modelos de propensión, y posteriormente se describe las que no pueden ser utilizadas según cada caso de estudio.

Sociodemográfico	Transaccional
Edad	Gasto en cada negocio
Cantidad hijos	Transacciones en cada negocio
Sexo	Cliente
Estado civil	Recency en cada negocio
Region	Crecimiento 3 meses vs 6 en cada negocio
Tarjeta	Share gasto web en cada negocio
Tenencia tarjeta	Email
Monto cupo	Envios (recibidos)
Tipo tarjeta	Aperturas
	Ratio de apertura

Tabla 5: Variables utilizadas

La totalidad de variables utilizadas da un total de 64, son principalmente variaciones de las que se resumen en la tabla anterior. Las variables de gasto y transacciones se separan por negocio y por canal en distintas agregaciones temporales (1, 3, 6 y 12 meses). La variable de región se separa por zonas regionales con tal de resolver el problema de contar con regiones con pocos clientes. El recency se refiere a cuantos días han pasado desde la última compra, donde se separa a nivel de holding y a nivel de Formato 1. Cliente corresponde al indicador de si es que el cliente cuenta con alguna transacción en el tiempo visto.

Las variables que no se pudieron utilizar para cada hito son las siguientes:

- Hito apertura de tarjeta: no se utilizaron las variables relativas a la tarjeta
- Hito primera compra web: no se utilizaron las variables transaccionales en el canal digital
- Hito cruce de negocio: En cada caso no se utilizaron las variables transaccionales correspondientes al negocio al cual se estaban cruzando.

7.1.3 LIMPIEZA DE DATOS Y TRATAMIENTO DE DATOS

En esta sección del trabajo se realiza una limpieza de valores atípicos (valores atípicos) utilizando herramientas visuales como boxplots y estadísticos comunes como las medias, medianas y distintos niveles de percentiles. Esto con la finalidad de observar grandes diferencias producidas por datos a típicos en la muestra. Posteriormente se establece un criterio general para todos los hitos con el cual se decide dejar fuera de cada una de las muestras los datos más extremos tanto en margen de contribución, gasto y contribución anterior.

Por otro lado, se revisa el poblamiento de las variables y se decide qué hacer con aquellas variables que no cuentan con el 100% del registro de los datos. Dependiendo del caso se puede excluir a los clientes de la base que no cuenten con datos en las respectivas variables, reemplazar con la moda, reemplazar con una categoría nula estilo: “Sin dato”, en caso de que sean variables categóricas o binarias. Asimismo, es posible reemplazar con valores arbitrarios pero que se intuye seguirán la misma lógica matemática (ejemplo de esto es el caso del recency, donde aquellos clientes que no han comprado en algún negocio en específico tendrán valor nulo para su recency, se puede rellenar su observación con un valor lo suficientemente grande para que el comportamiento de la variable en este cliente sea similar a clientes que no compran hace gran cantidad de días). A continuación, se muestra el análisis y limpieza para el caso del hito de primera compra web, las decisiones tomadas con respecto al tratamiento de datos son homologables para el resto de los hitos.

7.1.3.1 POBLAMIENTO DE VARIABLES

Como se resume en la tabla 6, se cuenta con aproximadamente 55.000 clientes activados y 800.000 clientes no activados antes de realizar la limpieza del hito de primera compra web.

Grupo	Clientes
No activado	~800,000
Activado	~50,000

Tabla 6: Base inicial de clientes, hito primera compra web

El primer paso corresponde a revisar que tan pobladas están las variables, donde se detecta que existe un número relevante de variables con un porcentaje importante de valores nulos. En la tabla siguiente se muestra el caso.

Variable	Porcentaje faltante
TIPO_TARJETA	63.25%
MONTO_CUPO	63.25%
ENVIOS	59.85%
APERTURAS	59.85%
OPEN_RATE	59.85%
REGION	14.54%
EDAD	11.96%
SEXO	0.01%

Tabla 7: Despoblamiento de variables primera compra web

Como se observa en la tabla 7, las variables tipo de tarjeta, monto cupo, envíos, aperturas y open rate cuentan con un gran nivel de despoblamiento. Como estas variables podrían representar información valiosa para el caso de estudio se decide no eliminarlas. Por otro lado, como es una gran cantidad de clientes los que no cuentan con datos para estas variables se decide no eliminar estas observaciones de la tabla. Finalmente, por el lado de la tenencia de tarjeta del holding se decide rellenar el tipo de tarjeta con el string “Sin tarjeta” y el monto cupo con 0. Por el lado de la apertura de mails, se decide rellenar los envíos, aperturas y el open rate con el valor 0. Se considera a priori que hacer estas agregaciones resulta beneficioso por el hecho de poder mantener tanto las variables y los clientes que no tienen valor en ellas, además se cree que no se estaría distorsionando la información entregada por las variables mismas.

Por el lado de región, edad y sexo, se decide mantener la variable y eliminar a los clientes que no tengan datos. Esto puesto que rellenar con la moda o con el promedio podría generar distorsiones en la información entregada por estas variables. Esto considerando que es un gran porcentaje de clientes los que no tienen valores en ellos y que, a diferencia de las variables del caso anterior, no existe ningún valor lógico para el caso de los nulos. Se entiende como lógico el hecho de que el monto cupo sea 0 cuando no se tiene tarjeta o el open rate sea 0 cuando no se reciben correos (o que el comportamiento de aquellos que no abren ningún mail también sea similar a los que no reciben ninguno, pues ambos no recibirán la información del holding).

Posteriormente a resolver el problema del despoblamiento de variables, se detecta que en el sexo de las personas se encuentra un caso a tratar. Existen clientes con sexo distinto a hombre o mujer como se muestra la tabla 8.

Sexo	Porcentaje del total (aproximado)
M o F	90%
Nulo	10%

Tabla 8: Sexo de los clientes primera compra web

Como se observa en la tabla 8, existe una cantidad importante de clientes que no tienen sexo F o M (las categorías estándar en sexualidad binaria). Este número representa cerca de un 10% de la muestra de clientes utilizados para la modelación de este hito. Reemplazar estos valores por la moda o repartirlos en igual cantidad no se considera correcto por el hecho de que podría estar agregando información incorrecta a las clases de hombre o mujer. Por otro lado, repartirlos mitad y mitad asumiría implícitamente que no existe un sesgo de sexo a la hora de registrar erróneamente este dato. Por otro lado, si se reparte aleatoriamente existe un 50% de probabilidad de categorizar erróneamente el sexo de la persona, o sea que en esperanza la mitad de las veces el sexo estaría errado. Este error podría neutralizar las diferencias de sexo en las estimaciones de efectos causales por lo que se decide no hacer esa opción. Finalmente, no se consideran a los clientes cuyos sexos no sean ni F ni M. Así, el saldo final de clientes para este hito, después de la limpieza de variables nulas es el que se muestra en la tabla 9.

Grupo	Antes	Después
No activado	~800,000	~650,000
Activado	~50,000	~50,000

Tabla 9: Limpieza de datos por poblamiento de variables, primera compra web

En la tabla 9, se observa una importante reducción en el número de clientes, donde se reduce la base inicial en un 17.4%. Resalta el hecho de que los clientes no activados fueron los principales eliminados. Esto se podría deber a que al mismo momento de realizar el hito (primera compra web) los clientes den los datos que faltaban en la limpieza anterior.

7.1.3.2 TRANSFORMACIÓN DE VARIABLES

Todas aquellas variables categóricas que tengan 2 opciones (como hombre o mujer) pasan a ser variables binarias de 1 o 0. Por otro lado también existen variables categóricas de más opciones, como es el caso de la región. Como existen algunas regiones que cuentan con un número pequeño de clientes, se decide agruparlas en macrozonas.

Zona	Porcentaje de Clientes
RM	42%
Centro Sur	26%
Centro Norte	14%
Norte	8%
Sur	9%

Tabla 10: Cantidad de clientes según zona, primera compra web.

Se puede observar que la región metropolitana es la más grande (tabla 10), pero que no representa más del 50% de los clientes. Las zonas céntricas, tanto norte como sur también albergan a una cantidad importante de clientes, más que las de los extremos sur o norte.

Por otro lado, para el caso de la edad, se decide generar una variable categórica según a cuál rango de edad pertenece el cliente, avanza de 5 años y queda constituida como sigue a continuación:

Rango Edad	Porcentaje clientes
18 a 23	9%
24 a 29	16%
30 a 35	15%
36 a 41	14%
42 a 47	14%
48 a 53	13%
54 a 59	11%
60 a 65	8%
66 +	10%
Sin identificar	0.02%

Tabla 11: cantidad de cliente según rango de edad primera compra web

El rango de edad más numeroso es el de 24 a 29 y 30 a 35 (tabla 11), los cuales van disminuyendo de forma constante hasta llegar a después de los 66 años. Se observa la

existencia de clientes con rango no identificado, estos son clientes que están en los rangos de 16 y 17, los cuales son eliminados posteriormente.

7.1.3.3 DETECCIÓN DE VALORES ATÍPICOS

El último paso del tratamiento y limpieza de datos es la detección y eliminación de valores atípicos. Las variables que se utilizan para detectar valores atípicos serán las de gasto en los últimos 12 meses (antes del hito) en los negocios F1, F2 y F3. Por otro lado, también se considerará el margen bruto de contribución en el holding sobre el gasto en el mismo periodo.

Como primer paso de detección de valores atípicos se calculan estadísticos comunes, como percentiles, medias, medianas y desviación estándar. Estos datos se pueden ver en la tabla 12.

Item	Gasto F1 12M	Gasto F2 12M	Gasto F3 12M	margen
mean	1.00	1.00	1.00	-1.00
std	2.04	2.22	2.19	450.40
min	0.00	0.00	-0.86	-327514.72
0.01	0.02	0.01	0.02	-1.37
0.1	0.07	0.04	0.07	0.12
0.25	0.16	0.11	0.15	0.37
0.5	0.42	0.36	0.40	0.60
0.75	1.18	1.05	1.01	0.80
0.9	2.57	2.38	2.29	0.98
0.99	7.39	9.28	9.59	1.28
max	736.57	306.07	240.56	396.77

Tabla 12: Estadísticos básicos de las variables de interés (normalizados según su media respectiva), caso primera compra web

De forma preliminar se sospecha la presencia de valores atípicos en todas las variables de la muestra mirando los datos de la tabla 12. Se cuenta con desviaciones estándares superiores al promedio de la variable, llegando a ser incluso el doble. Para el caso de margen la desviación estándar es de 450 veces la media. Por otro lado, tanto en Formato 1 como en Formato 2 se tiene que el valor mínimo de gasto es 0 veces la media, lo que no puede representar una compra real. Y el gasto en Formato 3 presenta un mínimo de gasto negativo, lo que, al no considerar devoluciones en el análisis, no puede estar correcto.

Por el lado del margen el mínimo es -327414 veces la media lo que no tiene sentido con la realidad. Resulta interesante que para esta variable el promedio sea negativo, lo que implicaría que el holding pierde dinero en promedio con las transacciones de sus clientes

(lo que claramente no es correcto). De esta misma forma entre el percentil 99 y el valor máximo de margen, existe una diferencia importante (absoluta), lo que podría implicar que existen valores atípicos entre el percentil 99 y el máximo. Sucede un caso similar con las variables de gasto, donde el percentil 99 ronda 7-9 veces la media, pero los valores máximos en algunas de las variables la superan 700 veces.

A continuación, se analizan boxplots de las distribuciones de las mismas variables anteriores para confirmar la presencia de valores atípicos. Estos se muestran en el gráfico 13, donde se puede apreciar que existen valores que se escapan de las observaciones más generales.

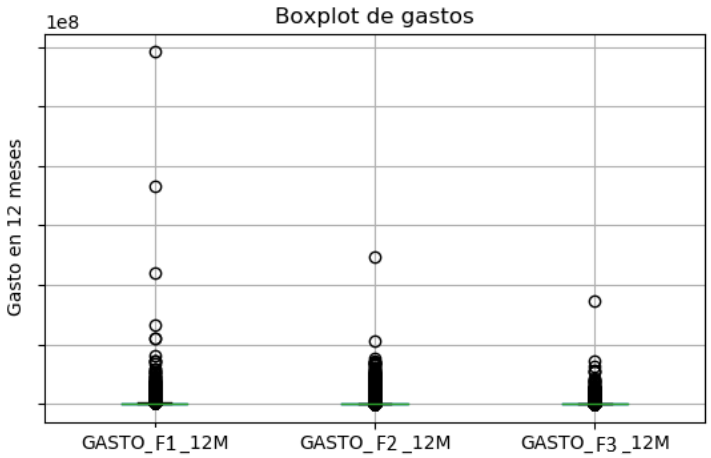


Gráfico 13: Boxplot de variables de gasto por negocio en 12 meses, caso primera compra web

En el gráfico 13, se observa como para todas las variables de gasto existe la presencia de valores anormales. En el caso de Formato 1 algunos datos que se encuentran muy alejados. Para el caso de Formato 2 esto también se hace evidente con la presencia de 1 de estos casos. Esto también sucede para los Formato 3 donde un solo cliente cuenta con compras millonarias. Así es como a causa de la alta presencia de datos atípicos es que los límites de los boxplot (25%, mediana y 75%) no son distinguibles a simple vista.

Gráfico 14: Boxplot de margen bruto, caso primera compra web.}

Un caso similar sucede con el margen, donde existe la presencia de valores que están sobre los 5 dígitos de margen negativo. Esto se puede observar en el gráfico 14.

Como se observa en el gráfico anterior, existe una observación que se encuentra aislada de las demás observaciones, el caso del mínimo. De todas formas, también existe un número no menor de observaciones con márgenes bajos. Estas observaciones son las que generan que el promedio del margen de toda la muestra sea negativo, pues se observa que, si bien existen algunos márgenes positivos grandes, los valores atípicos se concentran principalmente en el lado negativo del gráfico.

Finalmente se puede confirmar la presencia de valores atípicos en la muestra por lo que se procede a realizar una limpieza de datos. Para las variables de gasto, se pedirá que los clientes que cuenten con transacciones en el respectivo negocio, estas sumen por lo menos mil pesos durante la ventana de 12 meses. Por el lado de los valores máximos, se dejará fuera de la muestra a los valores superiores al 99.7% de la misma variable. Por otro lado, para el caso del margen se eliminará de la base a los clientes que pertenezcan al 0.3% inferior de los valores de margen y a los que sean superiores al 99.9%, esto dado que se observó que los valores atípicos están principalmente por el lado negativo de los datos. A continuación, se muestran los boxplots después de realizar la limpieza de valores atípicos.

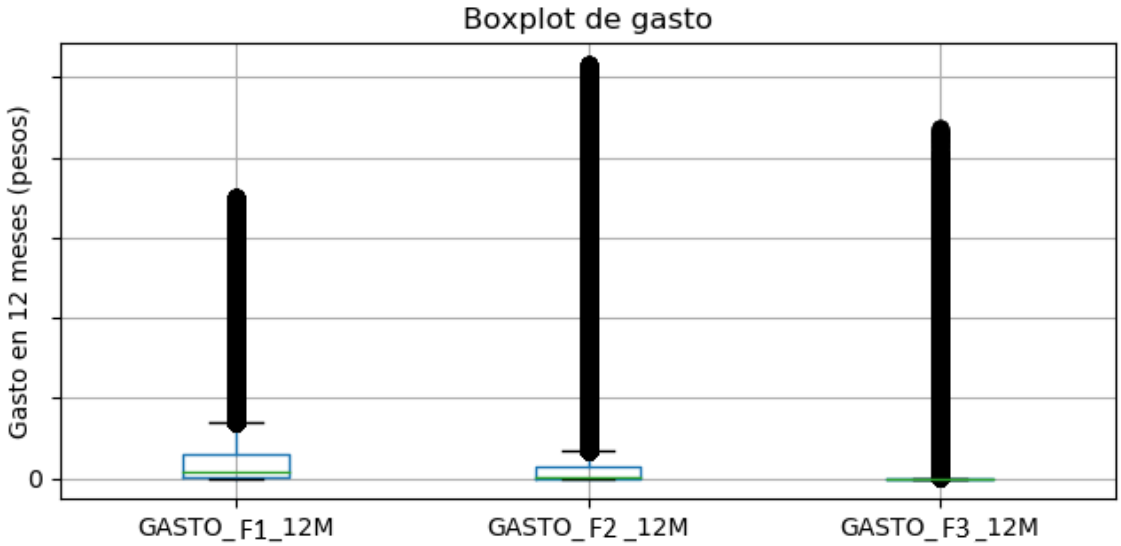


Gráfico 15: Boxplot de variables de gastos después de limpiar, caso primera compra web

En el gráfico 15, se puede ver los boxplot de las variables de gasto. Se ve como ahora los máximos están dentro de valores razonables. Si bien visualmente aun es complejo observar los límites de los boxplot, esto puede deberse a características propias del negocio, donde existe una gran cantidad de clientes que compra y no representan un gran porcentaje del gasto total percibido por los negocios, y por otro lado un número menor de clientes representa la mayor parte del gasto en los negocios.

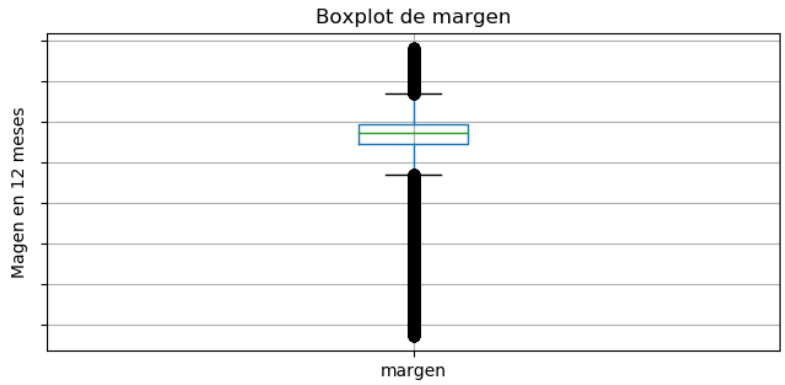


Gráfico 16: Boxplot de margen bruto después de limpiar, caso primera compra web.

En el gráfico 16, se pueden distinguir claramente los límites del boxplot de margen. Los valores negativos que escapaban de la distribución común y llegan a las cifras de miles han sido borrados. Aún persisten clientes con márgenes negativos, que implican pérdidas para el holding, pero estos no necesariamente son clientes anormales, son solo menos rentables. Cabe mencionar que ahora margen promedio si es positivo, lo que hace sentido para el caso de estudio.

En el anexo 2 se puede encontrar los estadísticos comunes que se mencionaron al comienzo de la limpieza de los datos, pero post limpieza. En general se observa que las desviaciones estándar de las variables disminuyeron, pero siguen siendo mayores a los promedios de las variables de gasto. Para el caso del margen el promedio ahora sí es positivo y la desviación estándar es menor al promedio. Por otro lado, las medianas siguen siendo menores a los promedios para el caso de las variables de gasto, lo que confirma la hipótesis de desigualdad en los patrones de consumo de los clientes, donde la mayoría de los clientes produce un porcentaje menor de las compras totales y existe un número menor de clientes que representa un gran porcentaje del gasto.

Para confirmar la presencia de desigualdad de gasto en las variables se hace el siguiente análisis: para cada negocio, qué porcentaje del total de clientes de ese mismo negocio representa el 80% de los ingresos percibidos. En la tabla número 13 se puede confirmar que el 80% del gasto es producido entre un 30-33% de los clientes en cada negocio, mientras el otro 20% es producido por el 67-70%. Cabe mencionar que esto es post limpieza de valores atípicos e imponiendo la condición de que sean clientes con al menos mil pesos de gasto si es que tienen 1 transacción en el negocio respectivo.

Negocio	80 % del gasto	20 % del gasto
TD	33%	67%
MH	30%	70%
SM	32%	68%

Tabla 13: Desigualdad de gasto, porcentaje de clientes que hacen 80 y 20% del gasto.

Finalmente, la eliminación de datos post limpieza de valores atípicos y la corrección del poblamiento de variables para el caso de primera compra web se resume según la ilustración 5.

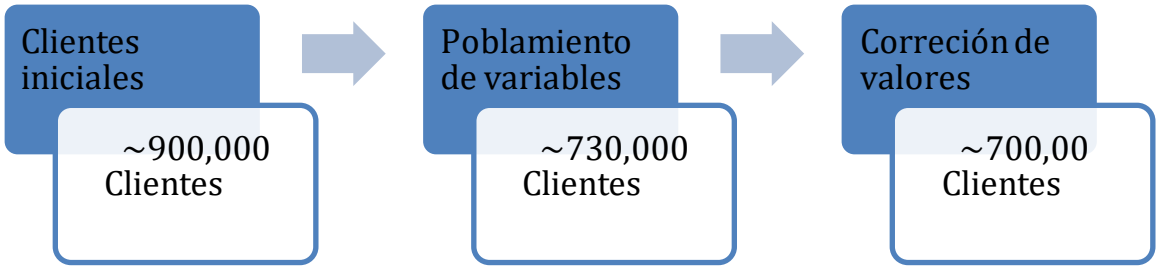


Ilustración 5: Proceso de limpieza y transformación de datos, caso primera compra web

Este proceso se repite para las bases de clientes de todos los hitos. La tabla 14 resume los resultados del proceso de limpieza de datos.

Hito	Grupo	Inicial	Post poblamiento	Post outliers	Eliminacion 1	Eliminacion 2
Compra web	Activado	824,789	674,947	666,507	18.2%	1.3%
	No activado	54,986	52,119	51,216	5.2%	1.7%
Apertura tarjeta	Activado	189,420	138,126	136,206	27.1%	1.4%
	No activado	12,628	12,610	12,454	0.1%	1.2%
TD a MH	Activado	1,488,689	1,147,033	1,130,353	23.0%	1.5%
	No activado	99,246	90,961	88,807	8.3%	2.4%
MH a TD	Activado	554,129	432,536	423,459	21.9%	2.1%
	No activado	36,942	33,195	32,488	10.1%	2.1%

Tabla 14: Resumen de cantidad de datos post proceso de limpieza, todos los casos (porcentajes con respecto a la cantidad inmediatamente anterior en el proceso)

Como se puede observar en la tabla 10, el resumen de limpieza de datos muestra que para todos los hitos el caso más importante de eliminación de observaciones es la corrección por poblamiento de las variables. Por otro lado, el proceso de eliminación de valores atípicos elimina un porcentaje mucho menor de las observaciones, donde el caso mayor es el del hito de cruce de negocio (desde Formato 1 a Formato 2), en que para el caso de los activados se elimina un 2.4% de valores atípicos.

7.1.4 ANÁLISIS DESCRIPTIVO

A continuación, se muestra el análisis descriptivo, donde se busca entender las distribuciones de gasto en los distintos negocios, ver si es que existen diferencias para las personas que realizan los hitos, y si es que estas varían en caso de género u otros variables.

Al igual que en la sección de limpieza de datos, en esta parte se ilustrará el análisis descriptivo del caso de primera compra web. Los gráficos más relevantes de los otros casos se pueden encontrar en anexos.

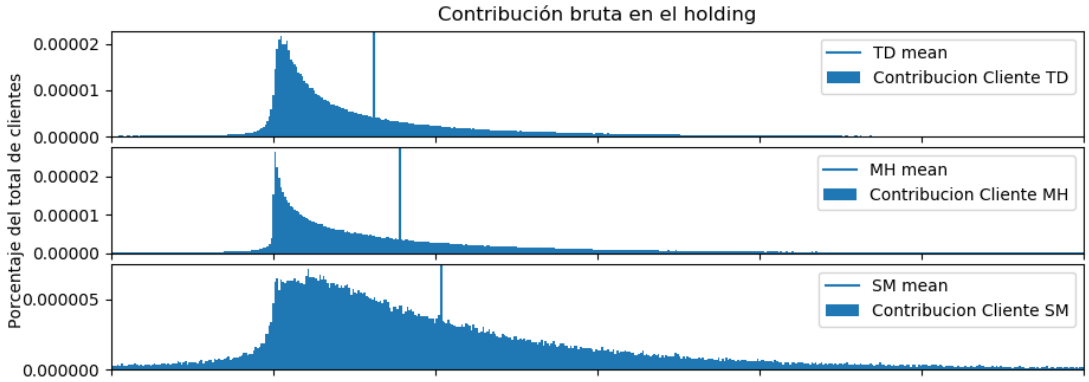


Gráfico 17: Contribuciones brutas del holding antes del hito

En el gráfico 17, se observa las diferencias en las distribuciones de contribución bruta del holding (antes del hito) compartiendo los ejes. Se puede apreciar un resultado medianamente contraintuitivo, los clientes con compra en Formato 1 son los que menos le contribuyen al holding, y los que cuentan con compras en los Formato 3 son los que más contribuyen. Esto no se condice con los datos de la introducción que muestran que la contribución bruta promedio del negocio de Formato 3 es la más baja del holding, superado por la de Formato 1 y Formato 2. Esto se produce principalmente por un problema de identificación, pues en su gran mayoría los clientes que cuentan con compras en Formato 3 son identificados solamente con la tarjeta del holding. A lo anterior también se suma que los clientes identificados en Formato 3 son en promedio clientes con compras en más negocios que en los otros. Estas hipótesis se pueden comprobar en la siguiente tabla.

Cliente del negocio	Clientes con tarjeta	Promedio de Negocios	Cantidad de clientes
Formato 1	46%	1.6	~600,000
Formato 2	56%	1.8	~400,000
Formato 3	90%	2.6	~100,000

Tabla 15: Descripción de los clientes de los negocios

Como se observa en la tabla 15, existe una diferencia importante en la proporción de clientes con tarjeta en cada negocio. Así, Formato 3 cuenta con un gran porcentaje de sus clientes que son identificados con tarjeta, a diferencia de Formato 1 que no supera el 50% y los de Formato 2 que superan levemente el porcentaje. Por otro lado, también resulta interesante que, en promedio, los clientes de SM son clientes en más negocios que los sus pares de Formato 2 y Formato 1. Una explicación posible de este fenómeno es que aquellos clientes que están fidelizados y cuentan con tarjeta del holding, tienen una mayor tasa de identificación en todos los negocios y los clientes que no están tan fidelizados, o no compran en los Formato 3 o tienden a identificarse con mayor medida en Formato 1 y Formato 2, pero no en Formato 3.

A continuación, se observa cómo cambian las distribuciones de gasto post realización del hito, separado para clientes de cada negocio. El gasto está medido sobre todo el holding.

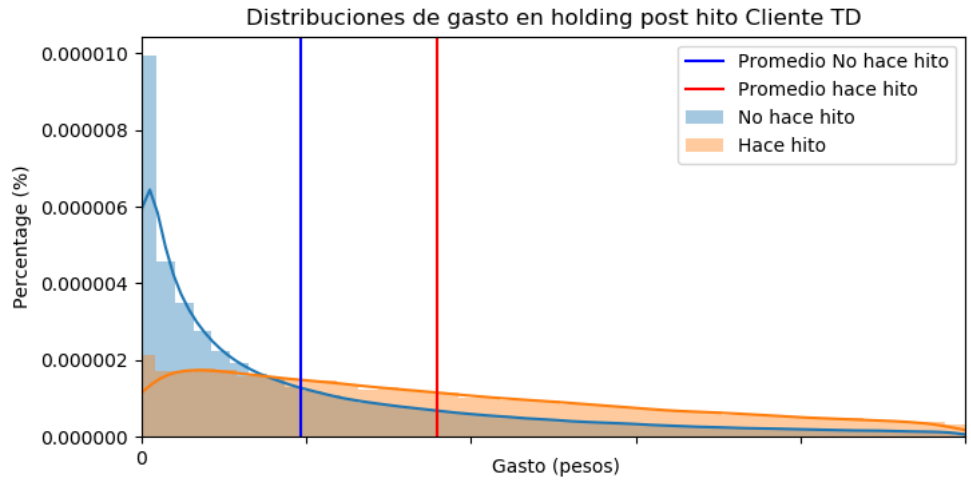


Gráfico 18: Gasto 12 meses en el holding para clientes de Formato 1, hito primera compra web

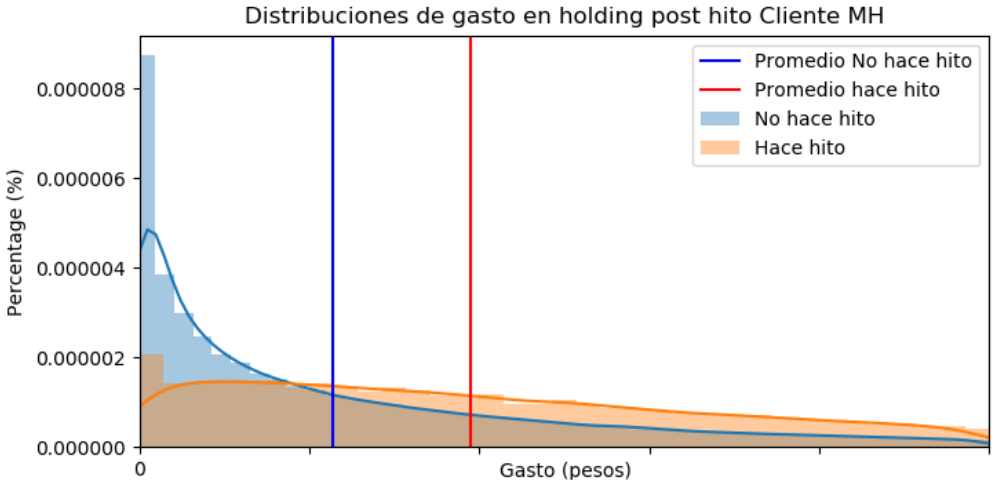


Gráfico 19: Gasto 12 meses en el holding para clientes de Formato 2, hito primera compra web

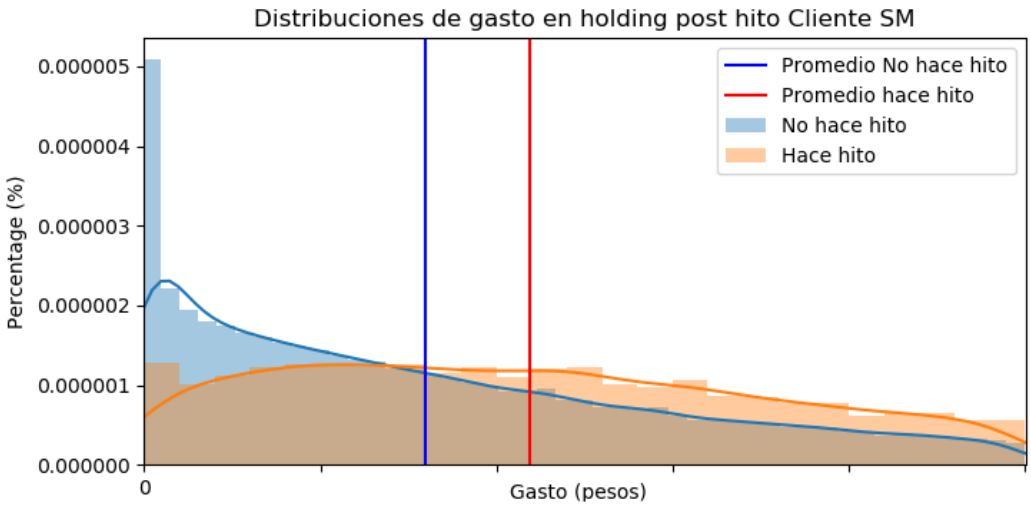


Gráfico 20 : Gasto 12 meses en todo el holding para clientes de Formato 3, hito primera compra web

En los gráficos 18, 19 y 20, se puede observar que la diferencia de los promedios de gasto en el holding para los clientes que realizan y no realizan los hitos es relevante para todos los tipos de clientes. En todos los casos se observa que la proporción de clientes con gasto 0 post realización del hito es menor para el grupo que realiza el hito.

Otro análisis interesante es ver si es que existen diferencias en las distribuciones de gasto para los clientes que realizan los hitos, pero previo a la realización de los mismos. La idea de este análisis es ver si es que las diferencias existentes en gasto holding antes de la realización del hito pueden explicar en parte las diferencias posteriores. En los gráficos 21, 22 y 23, se puede observar cómo eran las diferencias en las distribuciones de gasto de los clientes que en enero del 2017 realizan el hito y los que no.

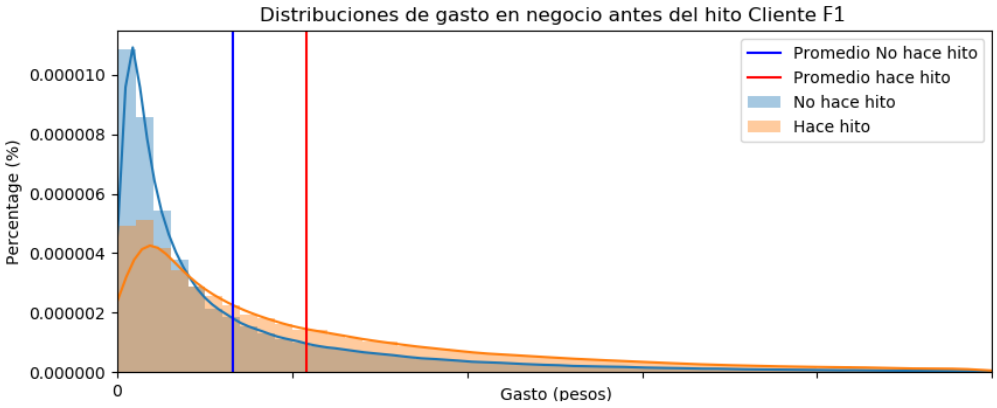


Gráfico 21: Gasto 12 meses en Formato 1, antes de realizar el hito primera compra web, grupos separados por realización de hitos

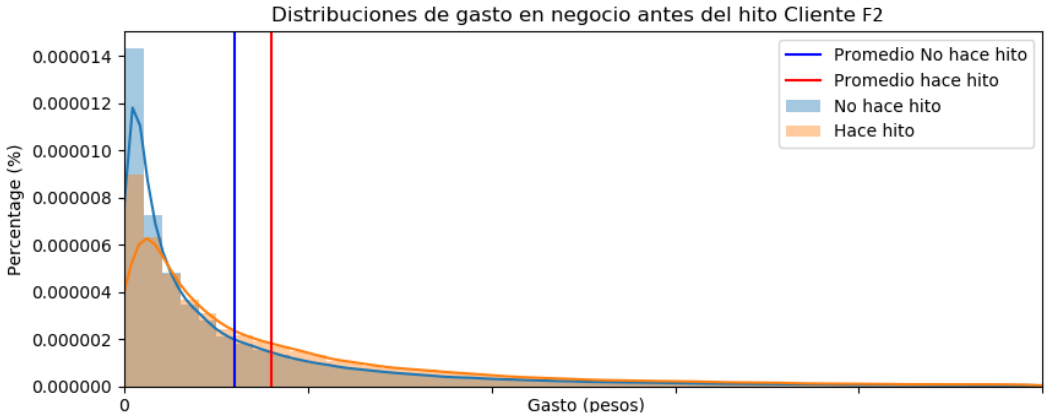


Gráfico 22: Gasto en negocio Formato 2, antes de realizar el hito primera compra web, grupos separados por realización de hitos

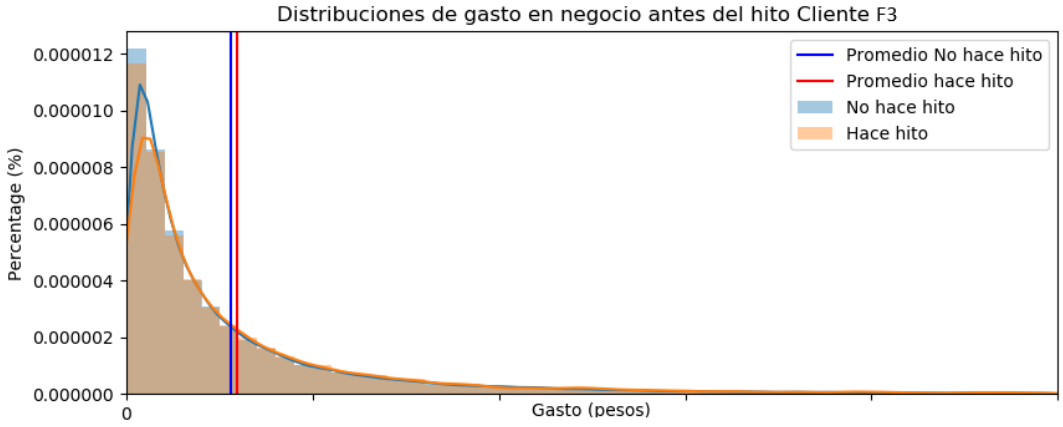


Gráfico 23: Gasto en negocio Formato 3, antes de realizar el hito primera compra web, grupos separados por realización de hitos

Del gráfico 21 se concluye que ambos grupos de clientes ya contaban con una diferencia en gasto importante en el negocio en el negocio Formato 1. Del gráfico 22 se puede concluir que los clientes de Formato 2 también contaban con una diferencia en el gasto anterior a la realización del hito, pero de menor magnitud que el caso de los de Formato 1. Sin embargo, en el gráfico 23 se aprecia que no existe una diferencia apreciable en el gasto anterior entre los grupos, por lo que probablemente no sea una variable relevante en la decisión de comprar por primera vez online en el holding.

Cabe mencionar que estos resultados de análisis descriptivo son solamente para el caso del hito de primera compra web y la base relacionada a la misma. Por lo que los resultados anteriormente mostrados no necesariamente se comportan de la misma forma para todos los hitos, puesto que las bases de clientes cumplen condiciones distintas.

7.2 ESTIMACIÓN DE EFECTOS

El primer ingrediente de la estimación de los efectos causales mediante PSM es calcular las probabilidades de los clientes de realizar los hitos justo antes de la ventana de observación de la realización de los hitos. Por lo que para poder hacer esto se ordenó la metodología de la siguiente forma, donde cada paso se describe a continuación:

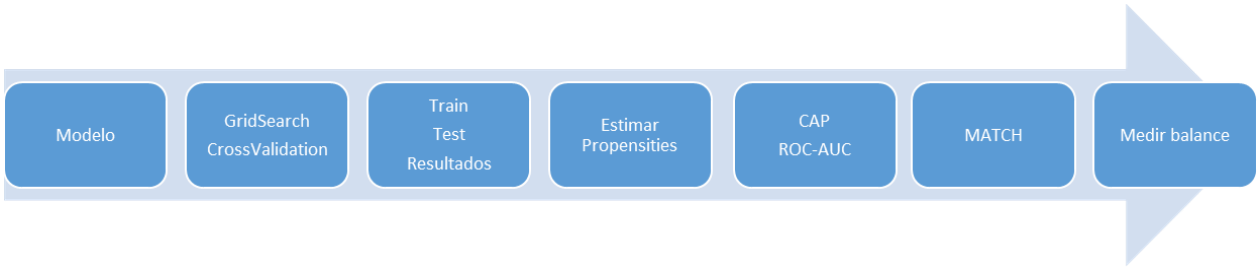


Ilustración 6: Metodología (primera parte) de la memoria

1) Modelo: Como se mencionó en el marco conceptual se cuentan con 3 modelos posibles para hacer el match. La decisión se tomará observando el balance post match y viendo cuál termina con muestras mejor balanceadas.

2) GridSearch y CrossValidation: para cada modelo se realiza un GridSearch con la finalidad de buscar aquellos hiperparámetros que impliquen un mejor desempeño de los clasificadores, y un CrossValidation para discriminar parámetros que puedan generar sobre ajuste en los datos.

3) Se observan los resultados para ver qué tan buen predictor es el modelo, además de ver la separabilidad de las muestras.

4) Se estiman los scores de propensión de los clientes mediante los modelos de clasificación calibrados

5) Se realiza el MATCH. Esto consiste en buscar de la lista de no activados al cliente que más se parezca en propensity score a cada cliente activado. Se hará un match de 1:1 y sin reemplazo pues es el que reduce el sesgo en la estimación de los efectos.

6) Después de realizar el match se ve qué tan balanceados quedan los grupos mediante la métrica de SMD. Se comparan el promedio de las SMD sobre las variables más importantes y también el máximo. En caso de tener algún máximo sobre el límite de 0.2 veces la desviación estándar del logit de los propensities se debe volver a estimar los propensity score, pero ahora agregando interacciones en las variables o nuevas formas funcionales (cuadrado, raíz cuadrada, etc.)

7) Se iteran los pasos anteriores variando el modelo a utilizar, se discrimina observando el modelo que resulta en el mejor balance.

8) Se estiman los efectos causales como promedio sobre toda la muestra de clientes

Posterior a la estimación de los efectos causales promedios se estiman efectos causales por segmentación. La segmentación se hace sobre el porcentaje de propensión, agrupando a los clientes en clusters según en qué rango de propensión de los hitos se encuentren. Como se menciona en el marco teórico son 3 métodos los que se probarán.

- Metodo de K-means con 5 clusters
- Quintiles de la muestra
- Cortes aleatorios de 5 clusters

La metodología detallada es la que se muestra en la ilustración

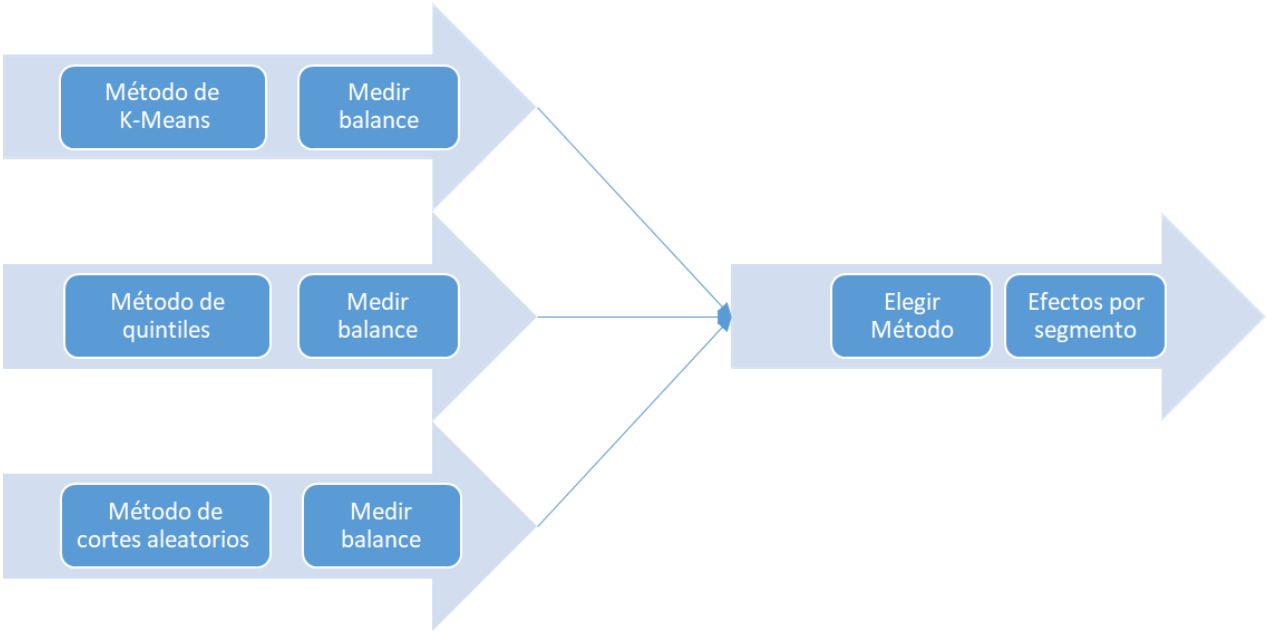


Ilustración 7: Resumen de metodología de estimación de efecto en segmentos.

En la ilustración 7, se muestra que de los 3 métodos de segmentación se elegirá el mejor en métricas de balance para estimar con ese método los efectos causales. La forma de discriminación es la misma que para el caso de los distintos modelos del punto anterior.

7.2.1 PRIMERA COMPRA WEB

7.2.1.1 GridSearch y CrossValidation modelos de propensión

Los hiperparámetros que se seleccionan acá serán los utilizados para calibrar nuevamente los modelos en la sección de los propensity scores. La principal finalidad de esta sección es encontrar los parámetros que muestren un menor nivel de sobre ajuste de los datos para el caso de los modelos más complejos, que usualmente tienden al sobre ajuste. Este es el caso de los modelos de Red Neuronal y Random Forest, que en ausencia de selección óptima de parámetros podemos obtener valores sobre ajustados, lo que sesga los propensity scores.

Los parámetros que se iteran son los siguientes para cada caso:

- Regresión logística:
  - Regularización l1 o l2: modificación de la función de costos que hace que el algoritmo que optimiza los costos de la regresión haga que los parámetros no lleguen a sus óptimos (para evitar sobre ajuste) o que elimine variables correlacionadas.
  - parámetro de costo de la regularización (C)
  - Intercepto: con o sin

- Random Forest:
  - Máxima profundidad de los árboles: una forma de evitar el sobre ajuste es limitar hasta donde pueden crecer las ramas de los árboles del bosque
  - Features máximos: cada split de cada árbol del bosque se hace mirando un conjunto máximo de variables las cuales puede elegir.

- Red neuronal:
  - Dropout: es una forma de evitar sobre ajuste, con una tasa decidida a priori se desactivan neuronas de la red cuando se calibra. Se varía la tasa con la que se desactivan neuronas
  - Número de capas: en este caso se tienen la opción de usar 1 o 2 capas en la estimación de la red, donde ambas capas son iguales
  - Número de neuronas: Se varía el número de neuronas de cada capa

Los parámetros finales que se seleccionan para estimar los propensitys son los siguientes para cada modelo:

Regresión logística: Regularización L1, C de 0.2335 y con intercepto
Random Forest: Profundidad de los árboles de 10 y features máximos de 10.
Red neuronal: Dropout de 15%, 1 capa y 27 neuronas.

Para la elección de los parámetros a utilizar, se calibró una única muestra de datos balanceada de tal forma que tuviera 50% de clientes activados y 50% de clientes no activados. Esto se hizo con la finalidad de poder comparar todos los modelos entre sí en métricas estándar de desempeño (Accuracy, Recall y Specificity), las cuales requieren un valor arbitrario para considerar a un cliente predicho como activado o no. En este caso se considera ese límite el 50%, por lo que un cliente con una probabilidad sobre 50% de realizar el hito es considerado como activado y bajo este número es considerado no activado (en las predicciones).

Por otro lado, si es que se utilizan otras razones de balance entre activados y no activados, como por ejemplo 1:4 o 1:9 (1 activado cada 4 no activados o 1 activado cada 9 no activados), se tiende a concentrar las distribuciones de las probabilidades cerca del 0. Por consiguiente, para poder comparar los desempeños de los modelos utilizando las métricas mencionadas, habría que mover el punto arbitrario del 50%. Considerando que la principal necesidad que suple este ejercicio es encontrar los parámetros que logran un menor sobre ajuste y mejor desempeño, no agrega valor hacer el análisis con otros niveles de balance ni otros puntos arbitrarios de corte.

Finalmente, los desempeños obtenidos por cada modelo de propensión son los siguientes con el umbral del 50% de probabilidad para declarar a un cliente como activado.

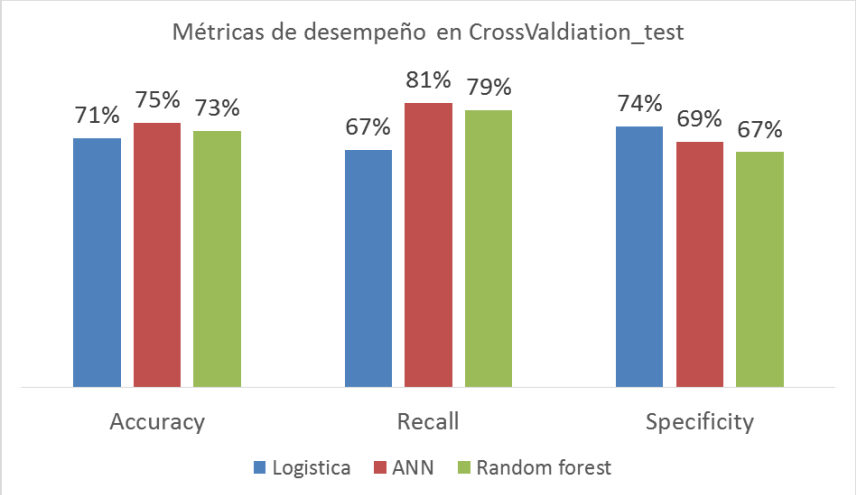


Gráfico 27: Métricas comunes de cada modelo en CrossValidation test set, caso primera compra web

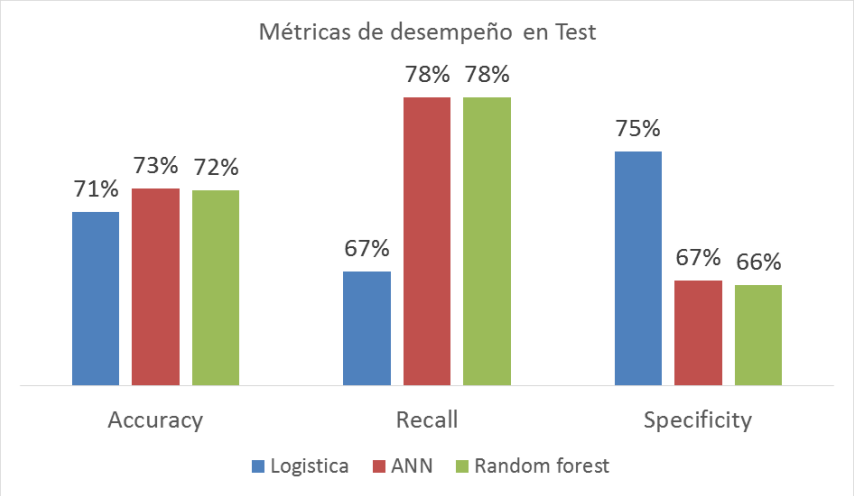


Gráfico 28: Métricas comunes de cada modelo en test set, caso primera compra web

Como se observa en los gráficos 27 y 28, tanto en accuracy como en recall los modelos de random forest y red neuronal tienden a presentar mejor desempeño. Pero por el lado del specificty tanto en los datos de test del CrossValidation y en el test final (hold out) se tiene que regresión logística supera a las otras. Esto probablemente se produzca por que el límite del 0.5 para el caso de la regresión logística es muy alto. También se presenta una diferencia entre estas 2 métricas para los modelos de Random Forest y Red Neuronal en el set de test, lo que evidencia que se podría nivelar estas métricas moviendo el criterio hacia el 1.

Así, se puede observar que los resultados son robustos al cambiar las muestras que se utilizan para la estimación de los modelos y la evaluación de los mismos. Por otro lado, en esta fase no se busca elegir el mejor modelo, se busca encontrar los mejores hiperparámetros.

7.2.1.2 Balanceo de clases

Para estimar los modelos de propensión la primera decisión que se debe tomar es cuál será el nivel de balanceo de clases con la que se calibrará los modelos para obtener después los propensitys sobre toda la muestra de clientes.

En esta sección se cuenta con 3 opciones:

- Balance de 1:1 (50% de activados y 50% de no activados), calibrar 10 clasificadores del mismo tipo con 10 muestras distintas de no activados y la misma de activados

- Balance 1:4 (20% de activados y 80% de no activados), calibrar 2 clasificadores del mismo tipo con 2 muestras distintas de no activados y la misma de activados

- Balance de 1:9 (10% de activados y 90% de no activados), calibrar 1 clasificador del mismo tipo con solo 1 muestra de no activados.

La razón para tomar 10 y 2 clasificadores distintos es lograr que los modelos aprendan de la mayor cantidad de datos posibles con la configuración de balance propuesta. Considerando que al tomar una muestra menor al tamaño real de los no activados se está perdiendo información que podría aportar a la estimación de los propensitys. Finalmente, la predicción final será el promedio de las predicciones de los modelos calibrados.

Para poder elegir el esquema a utilizar se toma en consideración 2 criterios:

Primero, el propensity score debe ser lo más cercano al real posible. Si es que se toma una muestra pequeña de los datos no activados y se calibran con el 100% de los activados, entonces se estará presente a pseudo valores de los propensities, los que podrían estar siendo sobredimensionados. Si se logra un buen balance a posteriori entonces, esto no será problema para la estimación de los efectos causales.

Sin embargo, contar con probabilidades no reales complica la propuesta de gestión. Esto porque la propuesta de gestión se basa en que se discrimine montos de inversión y priorización de clientes en base a rentabilidad esperada de la gestión de los hitos y no en base a propensión. Y en esta propuesta de gestión la probabilidad de realización basal de los hitos juega un papel crucial a la hora de discriminar que clientes gestionar en base al presupuesto. Por lo tanto, si se tiene valores de probabilidades no reales, y artificialmente más cercanas al 100%, hará que los clientes más propensos escalen rápidamente al primer lugar de la lista.

Segundo, se utilizará el criterio de balance de variables entre grupos (en SMD), donde mejor balance implicará menor diferencia de medias en las variables.

Por consiguiente, considerando estos 2 criterios es que se decide que a priori se toma un sampling menos drástico que el 1:1 siempre y cuando estos impliquen una mejora (o al menos no una perdida) en el balance de los grupos de activados y no activados en sus variables más importantes.

A continuación, y a modo de resumen se muestran los resultados de balance en variables cambiando los niveles de balance de muestras (1:1, 1:4 o 1:9) en todos los hitos. En base a este análisis se decide cual nivel de balance se utilizará en la estimación de todos los efectos causales.

Balances promedio	1:1		1:4		1:9	
Hito	Mean	Max	Mean	Max	Mean	Max
Primera Compra web	0.008	0.018	0.005	0.011	0.017	0.059
Apertura tarjeta	0.009	0.025	0.007	0.014	0.006	0.016
Cruce de F1 a F2	0.004	0.014	0.006	0.016	0.016	0.032
Cruce de F2 a F1	0.006	0.016	0.005	0.012	0.011	0.044

Tabla 16: SMD promedios y máximos sobre todo el grupo de activados y no activados después del match

En la tabla 16 se presentan los resultados del balance (medio en diferencia de medias estandarizadas) entre los grupos de activados y no activados después del match. Como se menciona en el marco teórico, el SMD es la principal métrica de desempeño con la cual se sabrá si es que un match es adecuado. El criterio para definir un match como bien hecho es que las variables más importantes medidas estén bajo el 0.1 de diferencia entre grupos (en valor absoluto).

Los resultados de la tabla 16 son los niveles de balance con el mejor modelo de las 3 opciones de clasificadores (Logística, Random Forest y Red Neuronal). Se puede observar que en general todos los resultados están bajo el 0.1 en promedio y los máximos de cada caso están también bajo el 0.1, por lo que en términos de balance de variables se podría usar cualquiera de las configuraciones de proporción de clases (1:1, 1:4 o 1:9) para la estimación de los efectos causales. Sin embargo, existe una tendencia a que los resultados de las configuraciones 1:1 y 1:4 sean similares y mejores en general que el 1:9.

Balances promedio	1:1		1:4		1:9	
Hito	Mean	Max	Mean	Max	Mean	Max
Primera Compra web	0.034	0.061	0.013	0.056	0.022	0.116
Apertura tarjeta	0.024	0.11	0.02	0.086	0.021	0.077
Cruce de F1 a F2	0.014	0.06	0.009	0.044	0.026	0.073
Cruce de F2 a F1	0.01	0.038	0.035	0.042	0.02	0.084

Tabla 17: SMD promedios y máximos sobre segmentos de grupo de activados y no activados después del match

En la tabla 17 se presentan resultados de balance de variables, pero en segmentación, por lo que el promedio mostrado corresponde al promedio de los promedios de los segmentos en cada caso, el método de segmentación utilizado fue cortes aleatorios. Por otro lado, el valor máximo representa al máximo de los máximos de los segmentos. En este caso, la única proporción de clases que nunca obtiene un valor sobre el 0.1 es el de 1:4, ya que tanto en 1:1 y en 1:9 se obtienen máximos sobre los 0.1 para sus segmentos.

Finalmente, viendo que los resultados en balance no son tan distintos para cada proporción de clases, se decide optar por utilizar 1:4 para la estimación de los efectos causales de todos los hitos. Esto pensando que, si bien los resultados son similares, existe una pequeña mejoría en balance en esa configuración, como también se cumple el criterio de que las clases tengan mayor proporcionalidad de no activados y así las propensiones estén más cercanas a las reales.

7.2.1.3 Propensión

A Continuación, se mostrarán los resultados de las métricas de desempeño para cada modelo de propensión.

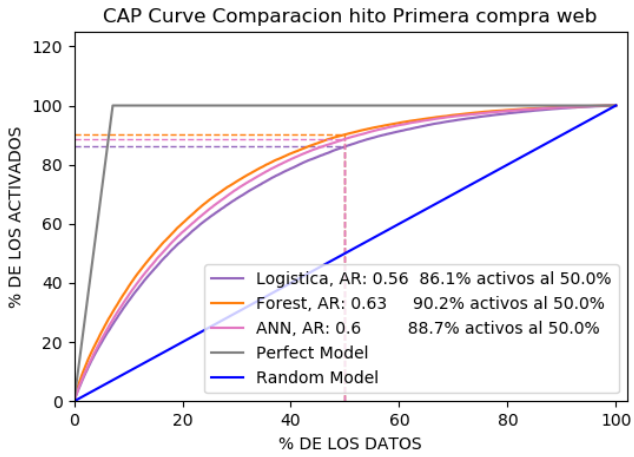


Gráfico 29: CAP distintos modelos primera compra web

Como se menciona en el detalle metodológico, como la selección de balanceo de clases fue 1 activado cada 4 no activados, entonces se calibran 2 modelos de cada tipo para después predecir sobre el 100% de los clientes. Esto logra los siguientes los resultados de curva CAP mostrados en el gráfico 29, se puede ver que el Random Forest es el mejor modelo para separar los grupos, obteniendo que un 90% de los designados como activados se logran detectar cuando se llega la mitad de la muestra ordenada por propensión. Los otros modelos cuentan con peores métricas en proporción de activados identificados al 50% al igual que área bajo la curva, siendo la regresión logística la que peor clasifica.

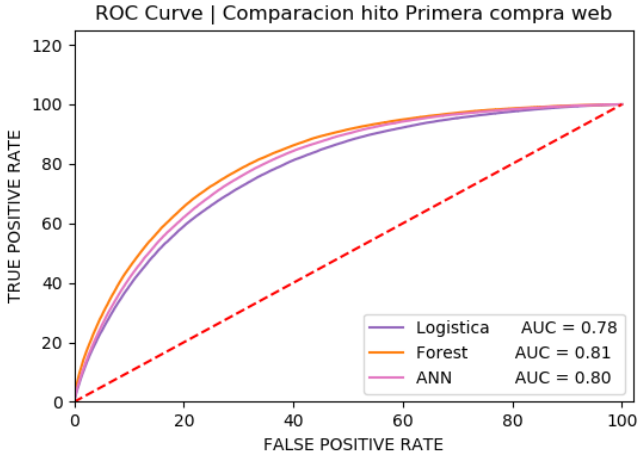


Gráfico 30: ROC distintos modelos para primera compra web

Por otro lado, también se cuenta con que la curva ROC-AUC (gráfico 30) confirma los resultados obtenidos por la curva CAP, donde el mejor clasificador es Random Forest, seguido por la red neuronal y en último lugar la regresión logística. Como se menciona en las secciones anteriores, estos gráficos son más robustos a la medida arbitraria de clasificación de un activado o no (límite del 50%). Dado que la curva CAP ordena a los clientes por propensión sin necesidad de elegir un criterio y la curva ROC hace el análisis de que tanto acierta a los activados y qué tanto le falla a los no activados para todos los valores posibles del criterio arbitrario en cuestión.

La finalidad de encontrar las propensiones de los clientes es poder posteriormente realizar el match emparejando a los clientes tratados y no tratados que más cerca estén en la probabilidad. Para avanzar en este fin, resulta de interés observar gráficamente la separación de las distribuciones de probabilidad de los clientes que se activan y los que no se activan. Lo que se puede observar en los gráficos 31, 32 y 33.

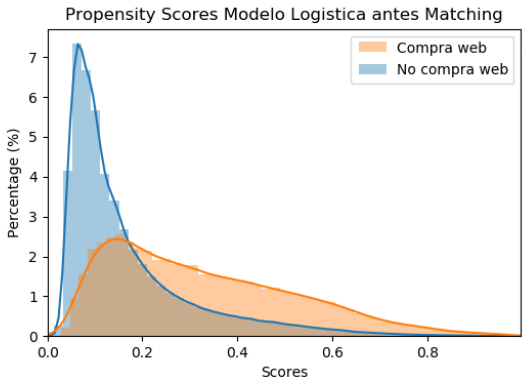


Gráfico 31: separación de clases con Logit

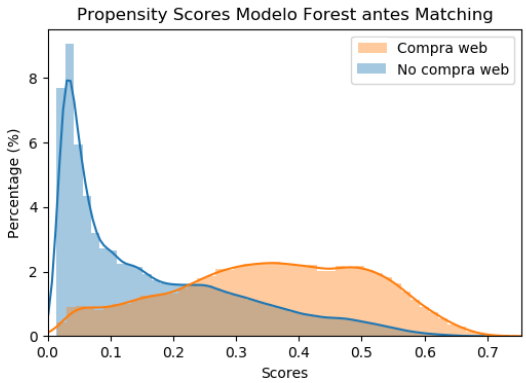


Gráfico 32: separación de clases con Random forest

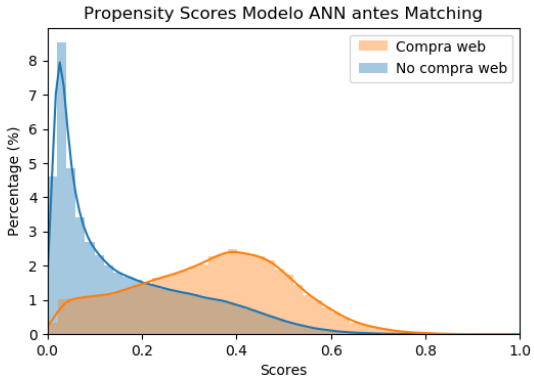


Gráfico 33: Separación de clases con Red Neuronal

En los gráficos 31, 32 y 33 se puede apreciar la separación de los grupos de activados y no activados en base a su propensity score calculado mediante cada uno de los modelos. En los 3 gráficos se puede apreciar una separación clara de las muestras, además que en todos ellos los puntajes de propensión están más cerca del 0 que del 1. Por otro lado, la distribución de probabilidad de los clientes que se activan tanto en el caso del modelo de Random Forest y Red Neuronal tiene un "peak" más cercano al lado derecho de su gráfico respectivo, esto es una diferencia importante con el caso de la distribución logística, lo que se puede deber a que estos modelos son más complejos y logran una mejor separación de los clientes. Por otro lado, una diferencia importante de los modelos es que el Random Forest tiene un máximo de probabilidad más bajo (cercano al 75%) que el caso de Logit y Red Neuronal, donde el máximo valor llega a un punto mayor.

El hecho de que todos los gráficos muestren visualmente una separación importante entre las clases es un punto a favor a la utilización del PSM como método de estimación de efectos causales. Puesto que esto representa una evidencia de que existen diferencias fundamentales entre aquellos clientes que se activan y no se activan, incluso después de haber hecho la limpieza de datos.

7.2.1.4 MATCH

Una vez calculados los puntajes de propensión con los 3 modelos de la metodología, el paso que sigue es la realización del match. Para cada cliente de la lista de activados se busca aquel cliente de la lista de no activados que tenga la menor diferencia de este puntaje con el activado en cuestión. Existe un criterio que tiene que cumplir esta diferencia de propensión para que se empareje satisfactoriamente al activado con el no activado: tiene que ser menor a un radio caliper definido previo al proceso de match. Este radio cumple el objetivo de desechar aquella dupla de clientes que, teniendo la menor diferencia en propensión posible, están aún alejados. En la literatura se encuentra que el radio óptimo para utilizar en PSM es 0.2 veces la desviación estándar conjunta del logit de los propensity scores (Austin P. C. 2010a).

De esta forma, los clientes activados que no tengan un cliente no activado a menor distancia que el radio caliper serán eliminados de la estimación del efecto causal. Entregando los siguientes resultados en términos de clientes satisfactoriamente emparejados.

Modelo	Match exitoso	porcentaje del total
Logit	51,042	99.7%
Random Forest	51,002	99.6%
Red Neuronal	51,014	99.6%

Tabla 18: Cantidad de parejas matcheadas exitosamente

Como se ve en la tabla 18, todos los modelos logran un nivel similar de clientes activados emparejados satisfactoriamente. El alto nivel de clientes matcheados es esperable pues el matching se está haciendo 1 a 1, o sea que existen varios candidatos para formar las parejas.

Como se mencionó anteriormente, en esta memoria se usa match sin reemplazo, esto significa que los clientes no activados no pueden ser utilizados para más de un match. Esto principalmente porque no se justifica repetir clientes en los emparejamientos si es que se tiene una proporción de activados y no activados relevante. En este caso la proporción es cerca de 10 no activados por activado, por lo que se decide no utilizar reemplazo.

7.2.1.5 VARIABLES MÁS IMPORTANTES

Para encontrar las variables más importantes en las cuales se medirá el balance entre grupos post match, se utilizan 2 métodos. El primer método consiste en un nuevo modelo de random forest con el cual se medirá la importancia relativa de cada variable, esta importancia se calcula midiendo cuanto ayuda cada variable a disminuir la métrica de impuridad de gini ponderada. Por otro lado, el segundo método será ver cuáles son las variables que en una regresión logística resultan ser significativas, y se eligen las primeras variables con mayor coeficiente en valor absoluto. Al estar todas las variables estandarizadas entonces los coeficientes son comparables.

Para el caso de las variables obtenidas por el bosque, se tomará aquellas variables que sumen en total el 70% de la importancia (ordenadas según importancia) y posteriormente por el lado de las de la regresión logística, se tomará el mismo número de variables que en el bosque. Finalmente, las variables importantes será la unión de los 2 grupos obtenidos mediante los dos mecanismos. Con esto se evita que el mecanismo utilizado para obtener las variables sesgue el resultado en beneficio del modelo de uno u otro modelo de propensión.

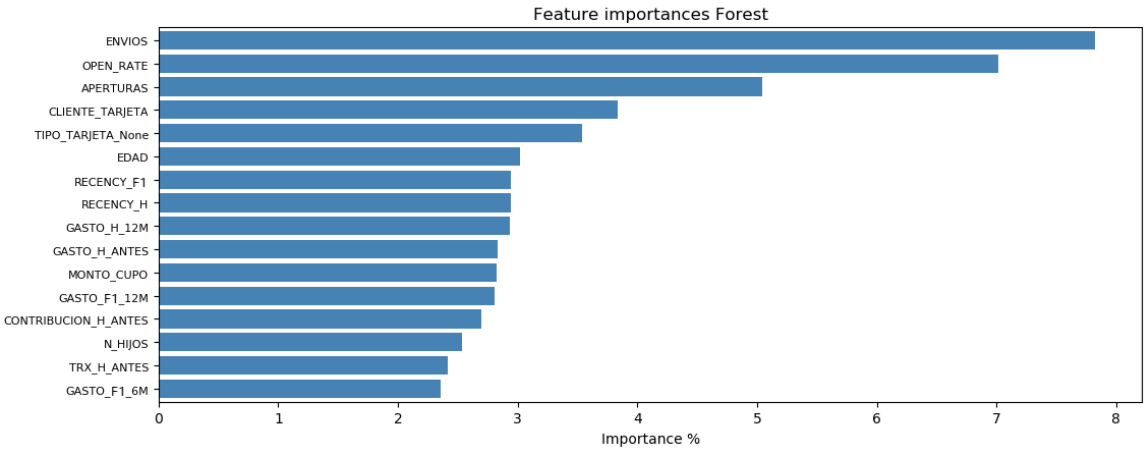


Gráfico 34: Importancia de variables según Random Forest, primera compra web

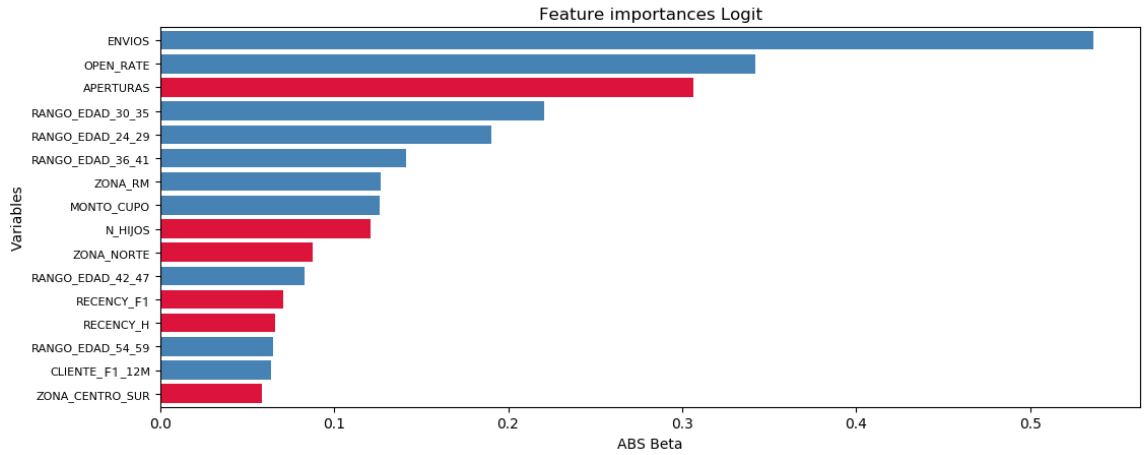


Gráfico 35: Importancia de variables según Regresión Logística, primera compra web. Rojo: valor negativo, azul: positivo.

Las variables más importantes según Random Forest (gráfico 34) son aquellas relacionadas con el canal digital como mails recibidos, tasa de apertura y cantidad de aperturas. Después vienen las variables relacionadas a la tarjeta como si es cliente de tarjeta y el tipo de tarjeta. Por otro lado, también están presentes las variables de gasto y recency para el negocio de Formato 1 y el nivel agregado por holding. En términos de sociodemográficos está la variable edad y la de número de hijos.

Por otro lado, en el caso de las variables más importantes según la regresión logística (gráfico 35) se repiten las variables relacionadas a los mails y de recency. Sin embargo, se agregan nuevas variables sociodemográficas, como la zona (correlación positiva ser de RM, negativo de Centro sur y norte). Interesante es que a diferencia de las importantes del Random Forest, el logit no considera edad por si sola como variable relevante, pero si algunos rangos de edad.

Finalmente, la unión de estas variables representa el set con el cual se prueba si es que los modelos generan un buen balance entre los grupos. El set unido representa un total de 25 variables de las 60 que se utilizan como input de los modelos.

7.2.1.6 Balance Post match

Una vez decididas las variables más importantes del caso de estudio, se debe medir el balance sobre la totalidad de estas variables para el match logrado con cada uno de los modelos.

modelo	SMD promedio	SMD maximo	SMD std
ANN	0.011	0.027	0.008
Logistica	0.018	0.073	0.022
Forest	0.056	0.119	0.034

Tabla 19: Resumen de balance en todas las variables

En la tabla 19 se puede apreciar el resumen de los resultados obtenidos. Como principal conclusión, todos los modelos obtienen balance promedio bajo el nivel de rechazo (0.1). Por otro lado, existen claras diferencias con respecto al Random Forest, que obtiene peores métricas de balance, si bien su nivel promedio se encuentra en 0.056, su nivel máximo esta sobre el 0.1, lo que significa que al menos una variable está desbalanceada con este modelo. Si bien no hay gran diferencia entre los promedios de la regresión logística y la red neuronal, la decisión será siempre aquel que logré un menor SMD promedio. Considerando lo anterior, se decide que la Red Neuronal es la que se utiliza.

A continuación, se desglosan los resultados de balance para las variables, donde se muestra la diferencia de medias estandarizadas antes del match y después, además del resultado de un t-test de medias donde la hipótesis nula corresponde a que las medias de los resultados son iguales.

VARIABLES	SMD antes	SMD después	P Value
GASTO_F1_12M	0.452	0.008	0.208
GASTO_H_12M	0.458	0.004	0.506
ZONA_NORTE	0.105	0.001	0.828
GASTO_F1_6M	0.404	0.009	0.165
MONTO_CUPO	0.509	0.017	0.007
ENVIOS	0.712	0.000	0.998
TRX_H_ANTES	0.423	0.003	0.630
GASTO_H_ANTES	0.458	0.004	0.506
CONTRIBUCION_H_ANTES	0.371	0.004	0.521
ZONA_RM	0.243	0.003	0.629
RECENCY_H	0.392	0.011	0.075
RANGO_EDAD_42_47	0.006	0.002	0.770
ZONA_CENTRO_SUR	0.131	0.002	0.770
EDAD	0.073	0.012	0.065
RANGO_EDAD_30_35	0.129	0.000	0.955
RANGO_EDAD_36_41	0.060	0.012	0.065
CLIENTE_TARJETA	0.661	0.019	0.002
N_HIJOS	0.149	0.000	0.938
APERTURAS	0.470	0.009	0.137
OPEN_RATE	0.550	0.001	0.921
RANGO_EDAD_24_29	0.073	0.005	0.458
TIPO_TARJETA_None	0.661	0.019	0.002
RANGO_EDAD_54_59	0.021	0.002	0.783
RECENCY_F1	0.426	0.018	0.003
CLIENTE_F2_12M	0.220	0.005	0.459

Tabla 20: Balance sobre las variables más importantes en ANN.

En la tabla 20 se observa el balance logrado por el modelo de Red Neuronal sobre las 25 variables más importantes de este caso. Antes de realizar el match la diferencia de medias estandarizadas (SMD) está sobre el límite de los 0.1 en 21 variables y solamente 4 variables se encuentra balanceadas (Los rangos de edad: 24 – 29, 42 – 47, 54 – 59 y la edad por si sola). Posterior a la realización del match, la diferencia de medias en la totalidad de las variables queda bajo el límite de rechazo, lo que significa que el match resulta satisfactorio.

Por otro lado, con la idea de analizar cuál sería el escenario si en vez de utilizar el SMD se utilizara un t-test de medias, se observa que en 4 variables se habría rechazado la igualdad de medias si se utiliza como margen el 0.05 de confianza. Estas variables justamente coinciden con aquellas de mayor SMD, pero este SMD sigue lejos del límite de rechazo.

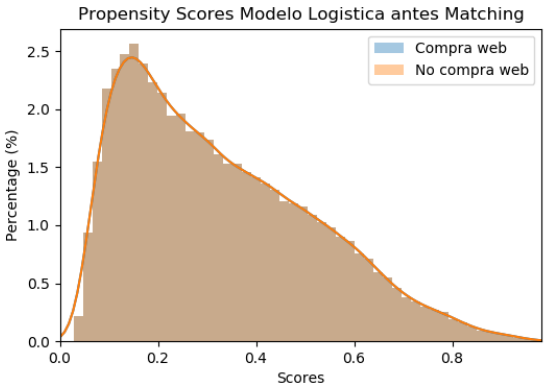


Gráfico 36: Distribución de propensitys con Logit

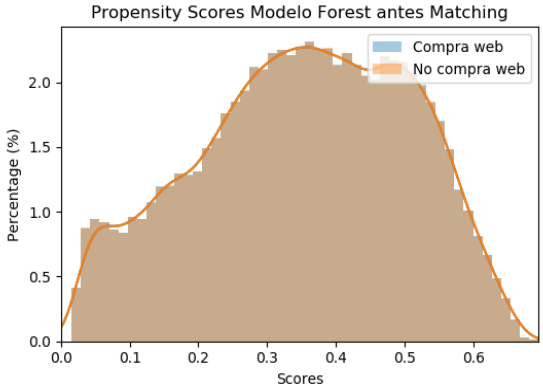


Gráfico 37: Distribución de propensitys con Forest

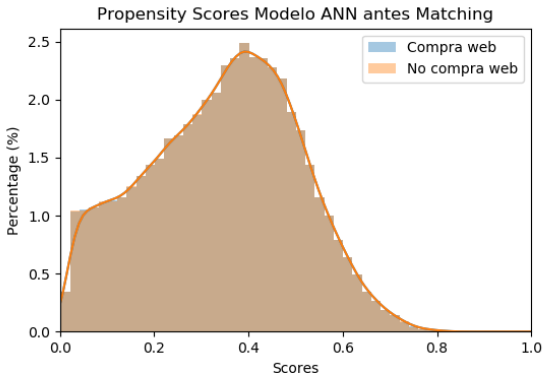


Gráfico 38: Distribución de propensitys con Red

En los gráficos 36, 37 y 38 se presenta la distribución de las propensiones de los clientes después del match con cada método distinto. Al menos en esta esta perspectiva se puede ver que todos los modelos generan distribuciones iguales y no distinguibles a la simple vista. Se puede observar que todos los modelos generan peaks y distribuciones distintas, lo que se debe a la forma en que los distintos algoritmos de aprendizaje de cada modelo resuelven el problema.

7.2.1.7 EFECTO CAUSAL PROMEDIO

Después de comprobar que los niveles de balance están sobre los mínimos exigibles para considerar el match como satisfactorio, entonces se puede pasar a la fase de comprobación de los efectos causales.

La estimación de los efectos causales es el efecto promedio sobre los tratados (ATT) medido como diferencia en diferencia. Esto es se calcula el incremento de la variable objetivo en cada grupo y después se restan estas diferencias, por lo tanto, la interpretación es: cuánto

más aumenta o disminuye el grupo de activados versus lo que aumenta o disminuye el grupo de no activados.

Las variables sobre las cuales se medirá el efecto causal son el gasto post match en 12 meses, la contribución en 12 meses y las transacciones realizadas por los clientes en 12 meses. Los cuales se resumen en la tabla número 21.

Variable	Grupo	Antes	Despues	Diferencia	Diff in diff
Gasto	Activado	1.00	1.51	0.50	0.28
	No activado	1.00	1.22	0.22	
Contribucion	Activado	0.99	1.35	0.35	0.23
	No activado	1.00	1.13	0.13	
Transacciones	Activado	1.00	1.37	0.36	0.19
	No activado	1.00	1.17	0.17	

Tabla 21: Efecto causal promedio primera compra web (normalizado al promedio de la variable del grupo no activado)

Todos los resultados y promedios se normalizan al promedio del grupo no activado en la variable respectiva. Esto se hace con la finalidad de mantener confidencialidad en los resultados sin perder riqueza en la información

Como se muestra en la tabla 21, el efecto en gasto es 0.28 veces el promedio de los no activados antes del hito. En promedio ambos grupos parten desde un nivel gasto cercano durante el año anterior, pero posterior a la ventana de observación del hito, los clientes que se activan aumentan en 0.5 su gasto versus los 0.22 de aumento de los clientes que no se activan.

Por el lado de la contribución, el efecto causal es de 0.23 promedios de los ni activados aproximadamente. Al igual que para el caso del gasto, ambos grupos parten desde un nivel similar, pero posterior al hito los activados aumentan 0.35 y los no activados aumentan solo 0.13.

Finalmente, con respecto a las transacciones, aquellos que clientes que se activaron en la ventana de observación obtuvieron un diferencial de 0.36 superior al diferencial de los no activados. Ambos grupos parten desde un nivel similar de transacciones.

7.2.1.8 MEJOR SEGMENTACIÓN

Posterior a la estimación de los efectos promedios, se puede avanzar a la estimación de efectos en segmentos de propensión. Como se mencionó en el marco conceptual, se utilizan

3 métodos de segmentación. El primero es mediante el algoritmo K-means basado en la métrica de propensión, el segundo es separar a los clientes en quintiles (20% de la muestra cada segmento) y el último es generar cortes aleatorios en la lista de clientes ordenados por propensión (con la restricción de que los segmentos resultantes no sean más pequeños que el 10% de la muestra).

Por lo tanto, y en consistencia con el resto de la metodología, el método de segmentación que se utilizará será aquel que provea las mejores métricas de balance entre los grupos de cada segmento. A continuación, se muestran los resultados de este ejercicio.

Segmento	Kmeans	Quintiles	Random cuts
1	0.021	0.022	0.017
2	0.012	0.018	0.013
3	0.019	0.013	0.010
4	0.018	0.015	0.016
5	0.015	0.018	0.015
Promedio	0.017	0.017	0.014

Tabla 22: Balance en segmentos según métodos de segmentación.

En la tabla 22 se resumen los resultados de balance en los distintos segmentos mediante los distintos métodos. Se ve claramente que los cortes aleatorios generan el menor balance posible, y si bien los otros segmentos dan peor balance, estos no empeoran tanto. Teniendo en cuenta que la heurística es elegir el método con mejor balance, las pequeñas diferencias

producidas pueden dejar abierta la decisión a otros criterios que piensen más desde la perspectiva de la gestión.

Finalmente, los segmentos escogidos son los de corte aleatorio, los que se componen de la forma en que se muestra en la ilustración Número 8.

Ilustración 8: Representación gráfico de los segmentos, hito primera compra web

En la ilustración anterior se puede ver gráficamente como se componen los segmentos. La forma que tienen los segmentos que producen los cortes aleatorios tienen una característica que resalta a la vista, tanto el segmento de menor propensión como el segmento de mayor propensión son los más extensos en términos de porcentaje. Por otro lado, los 3 segmentos de al medio parecen ser similares. Esto se puede confirmar con lo que se expresa en la tabla 24, donde se resumen las principales características de los segmentos.

Segmento	Cantidad de parejas	Limite izquierdo	Limite derecho	% del total
1	20,790	0.3%	31.3%	41%
2	5,829	31.3%	36.7%	11%
3	7,374	36.7%	42.9%	14%
4	5,641	42.9%	48.0%	11%
5	11,388	48.0%	91.3%	22%

Tabla 23: Composición de segmentos, primera compra web

Efectivamente, la tabla 23 confirma que el primer y el segundo segmento, no solo corresponden a los segmentos de intervalo de propensión más grande, sino que también corresponden a los segmentos con mayor cantidad de clientes dentro, con 41% y 22%. Los segmentos del medio son similares en porcentaje total de la muestra con la que cuentan, que es bastante cercano al límite de 10% exigido al método. Estas características representan una diferencia evidente con el método de segmentación de quintiles, pues este último contaría con 5 segmentos de igual cantidad de clientes en cada uno. Por lo tanto, los

cortes aleatorios son capaces de encontrar ordenamientos que generan un mejor balance, pero a priori pareciera ser que encuentra segmentos de menor interpretabilidad.

Por otro lado, el hecho de que los segmentos tanto del principio como del final tengan una mayor cantidad de datos, pareciera indicar que le método de matching funciona mejor en los clientes de propensiones cercanas al promedio. Por lo tanto, también esto indica que los grupos de clientes que se encuentran más en los extremos probablemente no logren encontrar tan buenos emparejamientos como los cercanos a la mitad del intervalo, necesitado una mayor cantidad de parejas para lograr un buen balance.

7.2.1.9 EFECTO CAUSAL EN SEGMENTACIÓN

Ya habiendo construido los segmentos y comprobado que el balance en los mismos es satisfactorio, se puede avanzar al objetivo de la segmentación, encontrar efectos causales heterogéneos en la muestra. Esto se realiza al igual que con el efecto promedio, en términos de diferencia y diferencia, pero ahora estimando estos promedios en cada uno de los segmentos formados. A continuación, se muestra un resumen gráfico de los efectos causales.

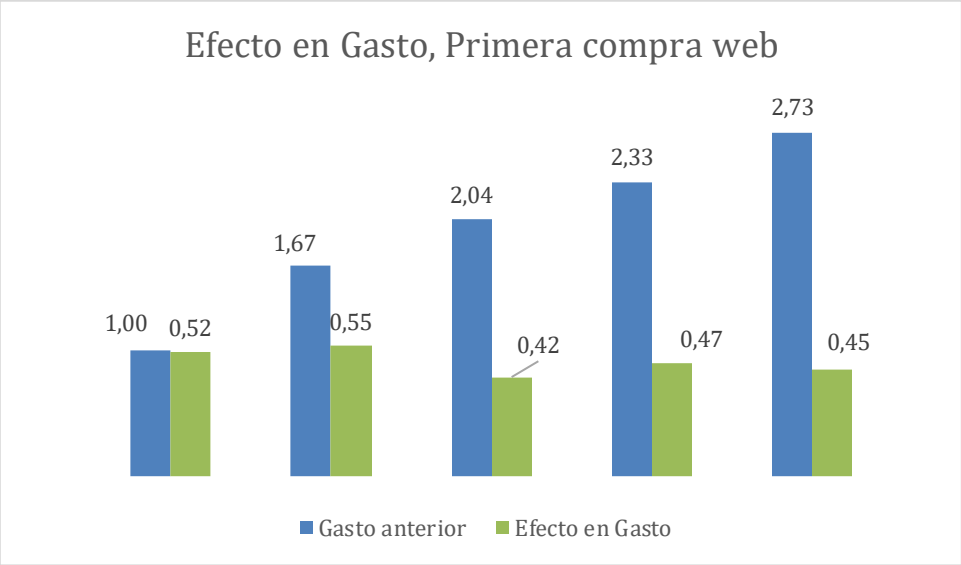


Gráfico 39: Efecto causal en gasto según promedio de gasto en segmento antes del hito (datos normalizados al gasto anterior del segmento 1)

En el gráfico 39 se muestra el efecto causal promedio (todos los datos normalizados al gasto anterior del segmento 1) en cada segmento y el gasto promedio durante un año antes de la observación del hito. En este gráfico lo primero que se observa es que en general, los segmentos a la vez que avanzan en propensión promedio a realizar los hitos también aumentan en gasto anterior a la realización del hito. Esto se debe que el gasto anterior es una variable que influye en el propensity. Por otro lado, se observa que no existe una tendencia tan clara sobre los efectos causales según se pertenece a un segmento de mayor

gasto. Esto significa que un cliente de mayor gasto no tendrá a su vez un efecto causal mayor a los clientes que pertenecen a segmentos de menor gasto. De hecho, se puede ver una leve disminución del efecto para los segmentos 3, 4 y 5 que son segmentos de gasto promedio 2.04, 2.33 y 2.73 veces el gasto del segmento primero.

Con respecto al efecto causal en segmentos para el caso de la contribución, se puede observar el gráfico 40. A diferencia del efecto en gasto, el efecto en contribución tiene una tendencia clara al momento de variar los segmentos. A medida que se avanza a un segmento con mayor contribución anterior, entonces se tendrá un efecto cada vez menor. Donde el segmento de gasto más bajo tiene un efecto de 0.53 veces el promedio de la contribución del primer segmento versus los 0.27 del segmento más alto, aproximadamente la mitad del efecto del primero.

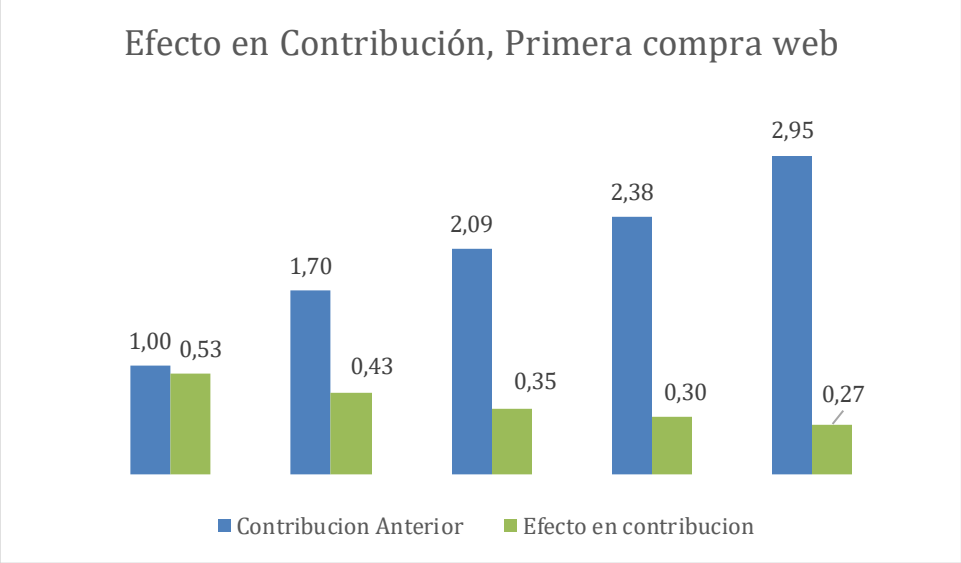


Gráfico 40: Efecto causal en contribución según promedio de contribución en segmento antes del hito (datos normalizados a la contribución anterior del segmento 1)

Una posible explicación de esto es que el efecto causal de gasto en estos mismos segmentos vaya disminuyendo, por lo que la contribución también tendría que disminuir. Sin embargo, como se ve en el resultado de gasto, la disminución de efecto en los segmentos mayores es leve y no llega a tener la baja que tiene la contribución. Por lo que esta opción es poco probable. Otra alternativa es que los clientes de mayor gasto sean clientes que buscan más exhaustivamente descuentos o que tiendan a comprar productos de menor contribución, si este es el caso entonces podría explicar porque el efecto para estos segmentos tiende a disminuir. Para poder probar la anterior hipótesis, es que se verá como varían los promedios de algunas variables relevantes que describen a los segmentos, de esta forma se podrá comprobar la existencia de tendencias en sus

variables explicativas sociodemográficas y transaccionales. Esto se hace después de hablar acerca de los resultados en transacciones.

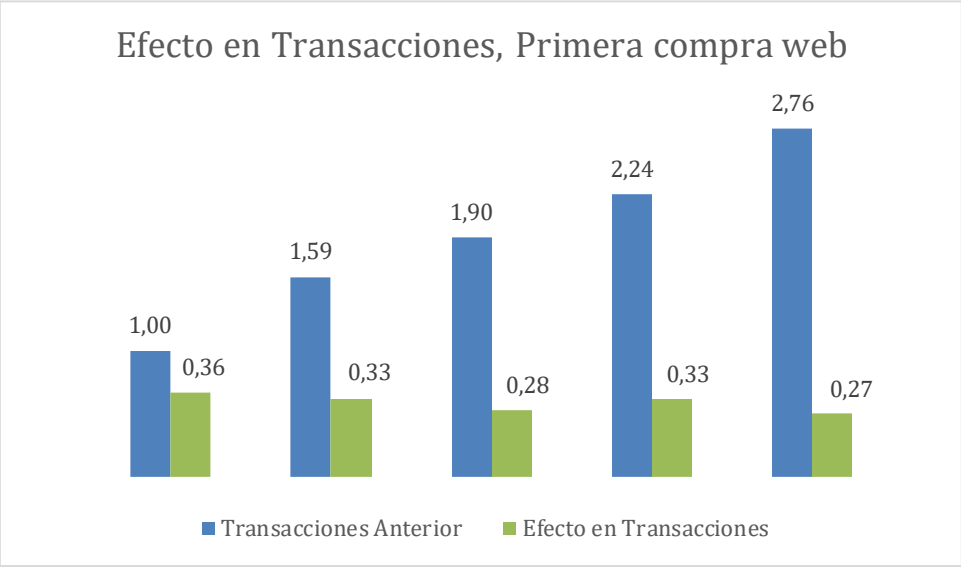


Gráfico 41: Efecto causal en transacciones según promedio de gasto en segmento antes del hito (datos normalizados a las transacciones anteriores del segmento 1)

El efecto causal en transacciones también muestra una tendencia a la baja a medida que se avanza en segmentos. Pero esta vez va más en concordancia con la bajada que se produce en gasto, la que no es tan alta como la baja en contribución. Finalmente se puede concluir que no porque los clientes tengan un gasto más alto, la realización del hito implicará un efecto mayor. En prácticamente todos los efectos se obtiene que disminuye con respecto al segmento de forma leve, excepto en la contribución que el efecto del último segmento representa la mitad del efecto del primer segmento.

Por otro lado, resulta interesante saber si es que estas disminuciones en los efectos causales con respecto al gasto se podrían deber a algunas diferencias en sus características sociodemográficas principales. Con este objetivo se plantea la tabla 25.

Variable	1	2	3	4	5
Edad	1.00	1.02	1.01	0.92	0.84
Hombre	1.00	0.96	0.96	0.93	0.78
Tarjeta	1.00	1.43	1.63	1.83	1.98
Hijos	1.00	1.01	0.98	0.84	0.57
Soltero	1.00	0.90	0.91	1.07	1.27
Zona norte	1.00	0.85	0.76	0.61	0.44
Zona centro norte	1.00	0.91	0.80	0.78	0.72
Zona RM	1.00	1.16	1.28	1.38	1.49

Zona Centro sur	1.00	0.84	0.75	0.68	0.64
Zona sur	1.00	0.96	0.86	0.77	0.60
Cliente TD	1.00	1.09	1.12	1.15	1.17
Cliente MH	1.00	1.17	1.26	1.32	1.36
Cliente SM	1.00	1.43	1.67	2.13	2.73

Tabla 24: Resumen sociodemográficos de los segmentos en primera compra web (normalizados al valor del segmento 1 en la respectiva variable)

En la tabla 24 se puede ver que los segmentos son distintos en los promedios de sus variables. Algo a resaltar es que los segmentos más altos en general son más jóvenes, tienen menos hijos y tienen una proporción más alta de personas solteras. Con respecto a la ubicación geográfica, a medida que se avanza en los segmentos se puede ver que disminuye la proporción de clientes en las zonas que no son Región Metropolitana, y esta última región justamente aumenta. Por otro lado, se produce un aumento importante en la proporción de clientes que compran en Formato 3. Esto último podría explicar la disminución en el efecto en contribución, pues los márgenes del negocio de Formato 3 son menores que el de Formato 1 o Formato 2.

7.2.1.10 INTERVALOS DE CONFIANZA

Un último resultado importante es obtener los intervalos de confianza de los resultados promedios con tal de ver que tanto podría variar el desempeño de una posible gestión. Para esto primero se muestran los histogramas de los efectos causales de cada cliente, esto es la diferencia de las diferencias de cada activado con su no activado y en segundo lugar los intervalos de confianza de los resultados promedios.

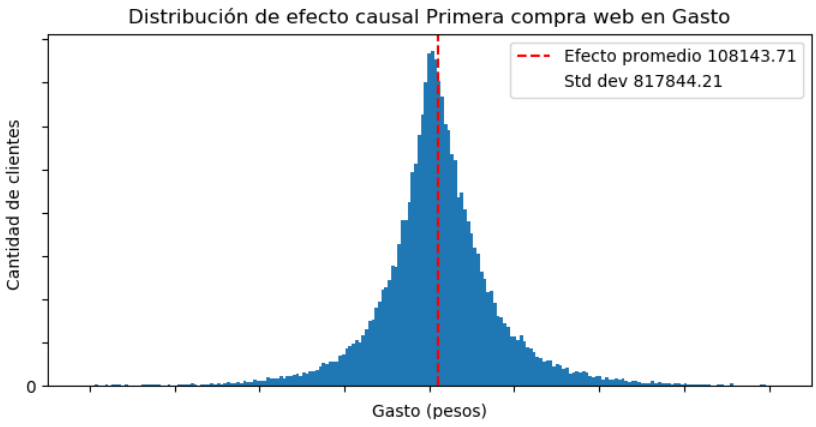


Gráfico 42: Distribución de efecto causal en gasto, primera compra web

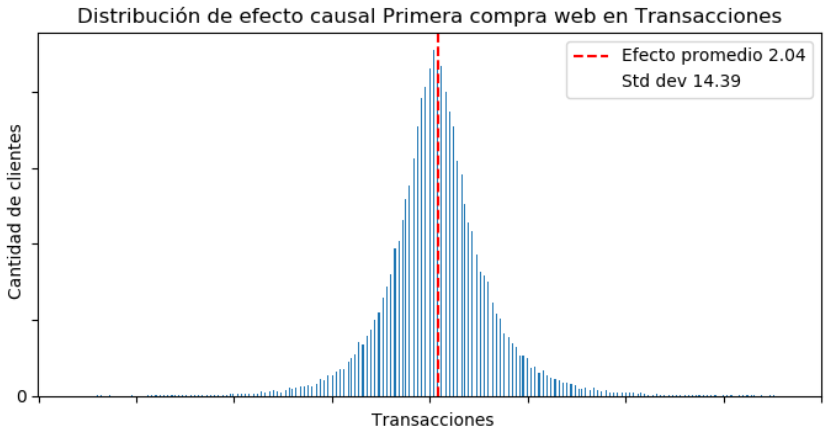


Gráfico 43: Distribución de efecto causal en contribución, primera compra web

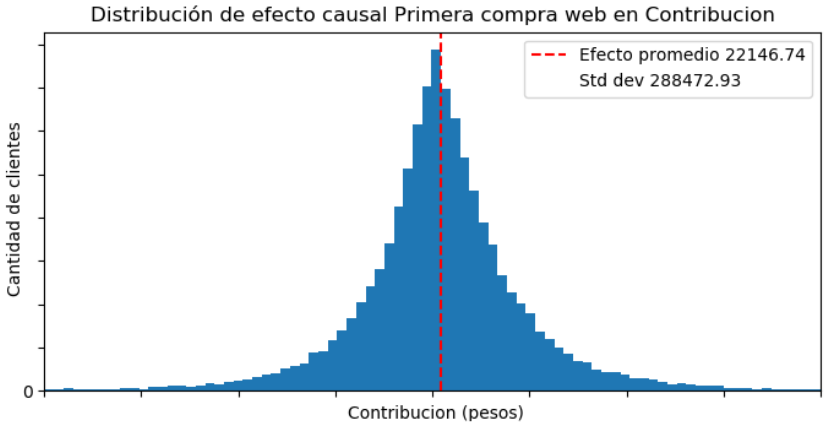


Gráfico 44: Distribución de efecto causal en transacciones, primera compra web

Como se observa en los gráficos 42, 43 y 44, el efecto unitario en cada cliente es bastante disperso, esto es esperable pues durante un año a un cliente le pasan más cosas que solamente el hito de la primera compra web. Pero lo que importa en este caso son los efectos promedios.

Se obtienen los intervalos de confianza al 95% y la significancia de los resultados mediante una regresión lineal en esquema de panel de datos.

Efecto	Promedio	Limite izquierdo	Limite derecho	Signficancia
Efecto Gasto	0.28	0.26	0.31	***
Efecto Contribucion	0.23	0.20	0.26	***
Efecto transacciones	0.19	0.17	0.21	***

Tabla 25: Intervalos de confianza y significancia para efectos promedios (datos normalizados al promedio de gasto, contribución y transacciones anteriores)
***: 99.9% de confianza

En la tabla 25 se resumen los intervalos de confianza para cada efecto causal. Cabe resaltar que con un 99.9% de confianza los resultados promedios de todos los efectos son distintos de 0 y con un 95% de confianza el valor de los efectos se encuentra en sus respectivos intervalos.

Gasto				
Segmento	Promedio	Limite izquierdo	Limite derecho	Signficancia
1	0.53	0.48	0.59	***
2	0.58	0.47	0.70	***
3	0.47	0.37	0.56	***
4	0.47	0.37	0.58	***
5	0.45	0.29	0.61	***
Contribucion				
Segmento	Promedio	Limite izquierdo	Limite derecho	Signficancia
1	0.53	0.45	0.62	***
2	0.49	0.35	0.64	***
3	0.35	0.23	0.47	***
4	0.32	0.19	0.45	***
5	0.27	0.07	0.47	***
Transacciones				
Segmento	Promedio	Limite izquierdo	Limite derecho	Signficancia
1	0.35	0.31	0.40	***
2	0.35	0.25	0.44	***
3	0.30	0.22	0.38	***
4	0.31	0.23	0.40	***
5	0.26	0.13	0.40	***

Tabla 26: Intervalos de confianza y significancia para efectos promedios en segmentos (normalizados al valor del gasto, contribución o transacciones anteriores del segmento 1)
***: 99.9% de confianza

Finalmente, los intervalos de confianza y significancia para los efectos promedios en segmentos se encuentran en la tabla 26, allí se puede observar que la significancia de todos los resultados es superior al 99.9%. Otro punto interesante es que en general en todos los efectos el intervalo de confianza del último segmento es relativamente más grande al de los demás, esto se puede deber a mayores niveles de desviación estándar en los efectos de ese segmento.

7.2.2 APERTURA DE TARJETA

A continuación, se muestran los principales resultados de la estimación de efectos causales para el caso de la apertura de tarjeta del holding. Los Boxplots para la limpieza de datos en el caso actual se pueden encontrar en el anexo 10.

Grupo	Clientes
0	~100,000
1	~10,000

Tabla 27: Cantidad de activados y no activados apertura de tarjeta

En la tabla 27 los clientes de grupo 1 son aquellos que sin contar con tarjeta del holding (cualquier tipo) durante diciembre del 2016, la abren en enero del 2017 y el grupo 0 son aquellos clientes que no realizan esta acción. Se puede observar que la cantidad de clientes que abrieron su tarjeta comercial durante enero del 2017 superan los 12 mil, se cuenta con una muestra de no activados de 136 mil.

Modelo	SMD promedio	SMD maximo	SMD std
Logistica	0.009	0.020	0.006
Forest	0.047	0.109	0.028
Red neuronal	0.018	0.037	0.009

Tabla 28: Comparación del balance según modelo

Para poder estimar los efectos causales con la metodología de PSM, primero se debe utilizar un modelo de propensión para obtener los scores de propensión sobre los cuales se realizará el matching. Por lo tanto, para decidir cuál modelo de propensión se utiliza, se realiza el mismo proceso que para el hito de la primera compra web. Así se obtienen los resultados de la tabla 28, donde se muestra que el menor SMD promedio se obtiene en el caso de la regresión logística, nuevamente los resultados no son tan distintos con los obtenidos mediante red neuronal, a diferencia del caso de Random Forest donde se obtiene un resultado inferior, más aun cuenta su máximo sobre el límite de rechazo (0.1).

Las variables más importantes (y sus valores de SMD) se pueden ver en los gráficos y la tabla del anexo 12.

Por consiguiente, el modelo utilizado es el de regresión logística. Por otro lado, en los gráficos 29 y 30 se pueden ver las principales métricas de desempeño de los modelos.

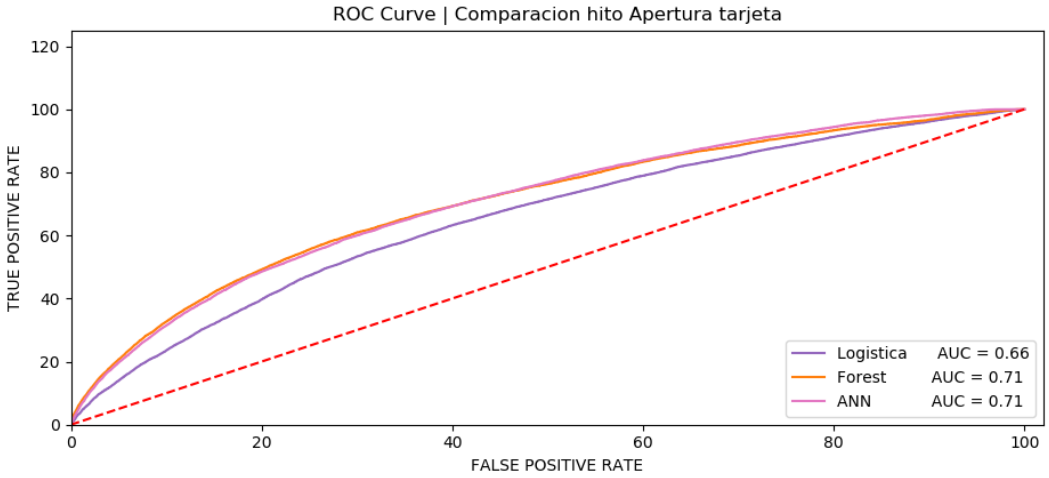


Tabla 29: ROC-AUC de los modelos de propensión, caso apertura tarjeta

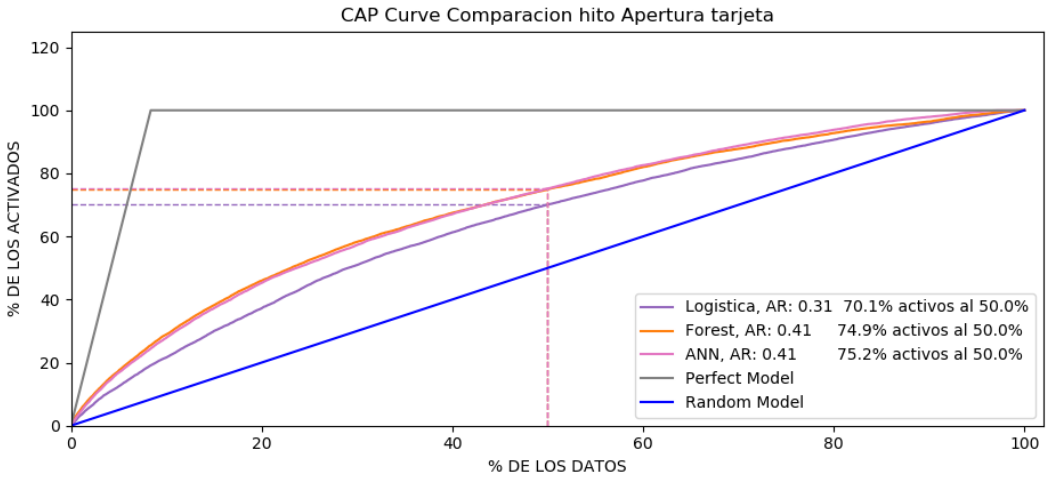


Tabla 30: Curva CAP de los modelos de propensión, caso apertura de tarjeta

De ambos gráficos de desempeño se concluye lo mismo que en la primera compra web, la regresión logística con un AUC de 0.66 y un AR de 0.31, tiene un peor desempeño a la hora de separar a los clientes activados de los no activados. Por otro lado, los modelos de Random Forest y Red Neuronal obtienen desempeños similares tanto en la curva ROC como en la CAP, con los mismos valores de AR y de AUC. A pesar de que Random Forest obtiene mejores métricas de clasificación, no logra superar a los otros 2 modelos

a la hora de ser utilizado como herramienta para el matching (en balance de variables medido en SMD). Por lo que utilizando el modelo de Regresión logística se obtiene la separación de muestras que se observa en el gráfico 45.

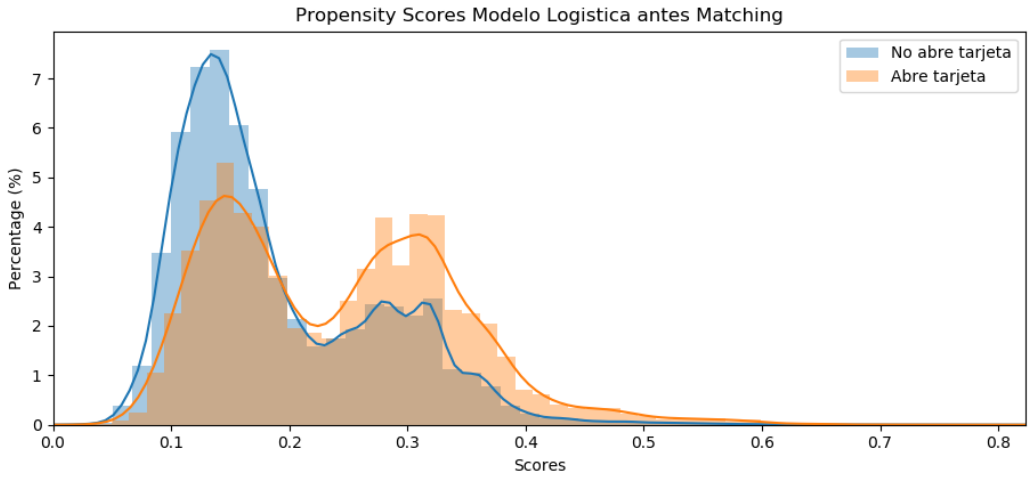


Gráfico 45: Separación de las distribuciones de propensión con logística, apertura de tarjeta

Así, habiendo elegido la regresión logística nuevamente como el modelo a utilizar para el matching, se obtienen los siguientes resultados de efectos causales en promedio.

Variable	Grupo	Antes	Despues	Diferencia	Diff in diff
Gasto	Activado	0.97	4.14	3.17	2.73
	No activado	1.00	1.44	0.44	
Contribucion	Activado	0.96	3.01	2.05	1.69
	No activado	1.00	1.36	0.36	
Transacciones	Activado	0.99	3.60	2.61	2.20
	No activado	1.00	1.41	0.41	

Tabla 31: Efectos causales promedios, caso de apertura de tarjeta comercial (normalizados al valor de la variable del grupo no activado)

A nivel promedio general, y según muestra la tabla 31. Los clientes de ambos grupos parten desde un nivel similar de gasto anterior. Posteriormente se produce una separación relevante en las diferenciales de gasto después del hito, pues los que se activan gastan 3.17 veces el promedio anterior y los que no solamente 0.44 veces más. Para el caso de la contribución se obtienen diferencias relativas similares, logrando un efecto superior 1.69 veces la contribución anterior. Por el lado de las transacciones,

aquellos clientes que no se activaron aumentaron en 0.41 en promedio y los activados lo hicieron en 2.61, nuevamente una diferencia relevante.

Las diferencias encontradas para el caso de la apertura de la tarjeta comercial son cerca del triple que las de la primera compra web. Esto puede ser explicado por varias razones, entre ellas, que los clientes con tarjeta comercial son mejor identificados pues al utilizar como medio de pago la nueva tarjeta sus compras son registradas automáticamente. También existe un mayor incentivo a identificarse a la hora de realizar compras, pues se acumulan puntos en el sistema de fidelización del holding, los que pueden ser canjeados por premios. No obstante, también existe el efecto causal producido por la obtención de la tarjeta, pues se cuentan con beneficios exclusivos a la hora de comprar en los negocios del holding, lo que se traduce en incentivos para traer gasto de la competencia a los negocios propios.

Una vez obtenidos los resultados promedios se procede a obtener los resultados en segmentos. Para esto se utiliza el método de cortes aleatorios, pues es el que genera el mejor balance entre segmentos. Obteniendo un nivel promedio de 0.024 de SMD, seguido en segundo lugar por el método de quintiles que genera un SMD de 0.027. Si bien la diferencia no es tan relevante, el criterio es seleccionar al que genere mejor SMD, por lo que se utiliza el método de cortes aleatorios. La tabla completa de los resultados en balance se puede ver en el anexo 15.

Segmento	Cantidad de parejas	Limite izquierdo	Limite derecho	% del total
1	2,688	3%	15%	22%
2	1,982	15%	18%	16%
3	2,560	18%	27%	21%
4	2,618	27%	32%	21%
5	2,606	32%	78%	21%

Tabla 32: Composición de segmentos caso apertura tarjeta comercial

En la tabla 32 se muestra como quedan compuestos los segmentos de este hito. Una diferencia notoria con el caso de primera compra web es que los segmentos quedan compuestos por cantidades similares de clientes, con 4 segmentos con levemente más del 20% de los clientes y uno con 16%. Por otro lado, se observa que este mismo segmento con menor cantidad de datos, tiene límites bastante cercanos en porcentaje de probabilidad, partiendo desde el 15% hasta el 18%. El último segmento es el que presenta

mayor dispersión en sus propensiones, abarcando más de 35% del dominio de probabilidad. Esto se puede ver gráficamente en la ilustración 9.

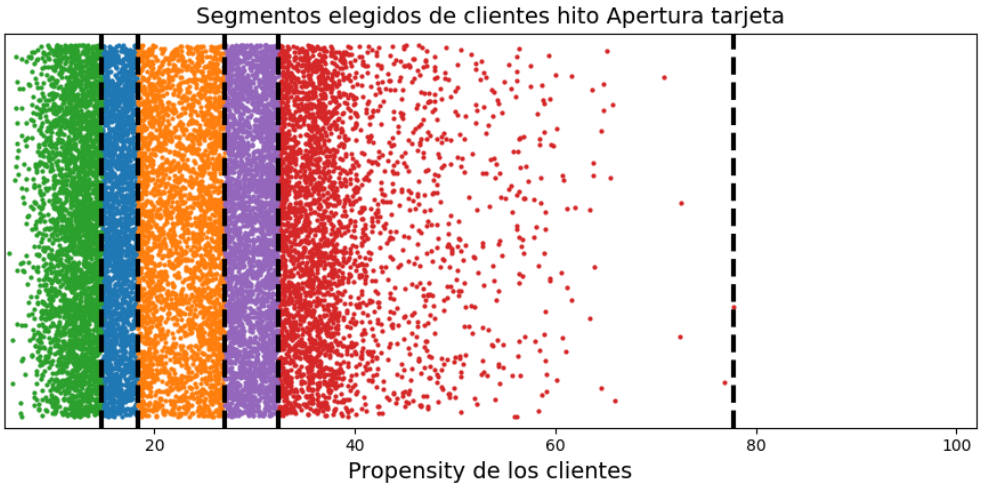


Ilustración 9: Composición de los segmentos de apertura de tarjeta.

Con la finalidad de entender la composición de clientes de cada uno de los segmentos, en la tabla 33 se puede observar los valores promedios de algunas de las variables sociodemográficas y transaccionales. Algo que resalta es que no hay grandes diferencias en los niveles de gasto anterior a la hora de definir la pertenencia de un cliente a un segmento, lo que se podría deber a que este tipo de tarjeta comercial apunta a segmentos de clase media (los cuales tienen baja dispersión de ingresos). Por otro lado, existe una tendencia clara a que los clientes más jóvenes (promedios van desde la edad de 46 hasta los 25 años) tiendan a tener porcentajes de propensión más altos y a su vez a pertenecer a segmentos del tipo 4 o 5. También, existe una tendencia en proporción de mujeres, personas solteras, de la región metropolitana y clientes de Formato 1 a medida que se avanza en los segmentos, lo que va dando un perfil bastante claro de cuáles son los clientes más propensos a abrir tarjeta.

Otro punto interesante es que existe una diferencia con el caso de la primera compra web, los segmentos más altos cuentan con una menor proporción de clientes de tiendas Formato 2 e invariabilidad en la de clientes de Formato 3. Finalmente, otra tendencia positiva y al igual que el caso de primera compra web, es que la propensión se correlaciona negativamente con la tenencia de hijos, bajando desde un promedio de 2.08 en el segmento 1 hasta 0.39 en el segmento 5.

Variable	1	2	3	4	5
Gasto anterior	1.00	1.08	1.33	0.73	1.39
Edad	1.00	0.93	0.75	0.55	0.55
Hombre	1.00	0.49	0.79	0.64	0.24
Hijos	1.00	0.81	0.49	0.23	0.19
Soltero	1.00	2.09	2.98	3.61	3.68

Zona norte	1.00	0.77	0.96	0.78	0.58
Zona RM	1.00	1.29	1.15	1.20	1.47
Zona Centro Sur	1.00	0.62	0.79	0.82	0.49
Zona sur	1.00	0.97	1.14	1.02	1.03
Cliente TD	1.00	1.39	1.31	1.48	1.51
Cliente MH	1.00	0.61	0.71	0.33	0.37
Cliente SM	1.00	0.55	0.74	0.23	0.62

Tabla 33: Valores promedios de variables de los clientes, apertura de tarjeta (normalizados al valor del segmento 1 en la respectiva variable)

Ya habiendo generado y descrito los segmentos se pueden obtener los resultados en cada uno de ellos. Al igual que en el hito anterior, se obtienen resultados para gasto, contribución y transacciones.

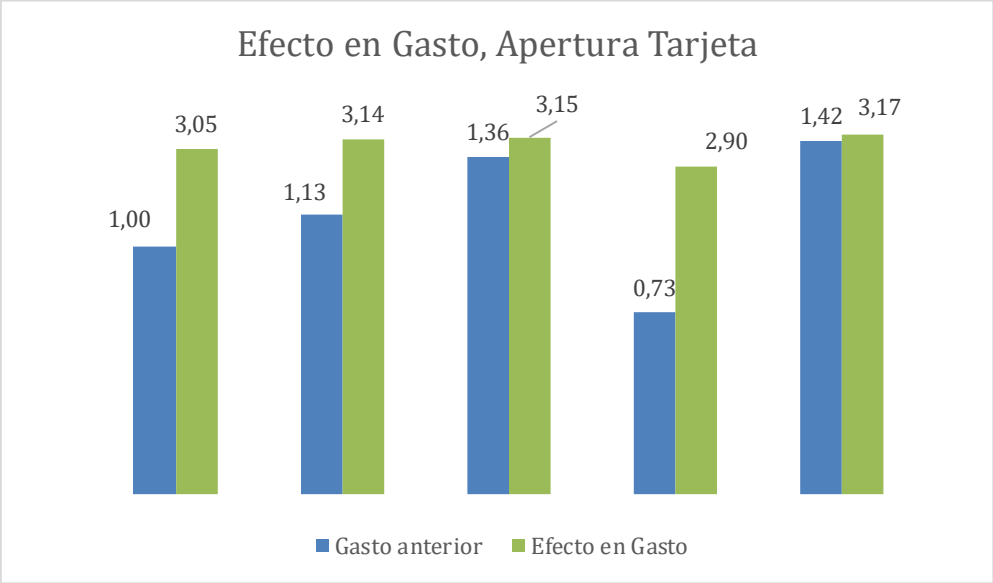


Gráfico 46: Efecto en gasto por segmentos, apertura tarjeta (datos normalizados al gasto anterior del segmento 1)

En el gráfico 46 se presentan los efectos causales por segmento para el gasto. Se puede apreciar que no existe ninguna tendencia al alza o a la baja en el efecto en medida que

se avanza en los segmentos, si no que se obtienen resultados variando entre los 0.52-0.42 veces el gasto anterior del segmento 1.

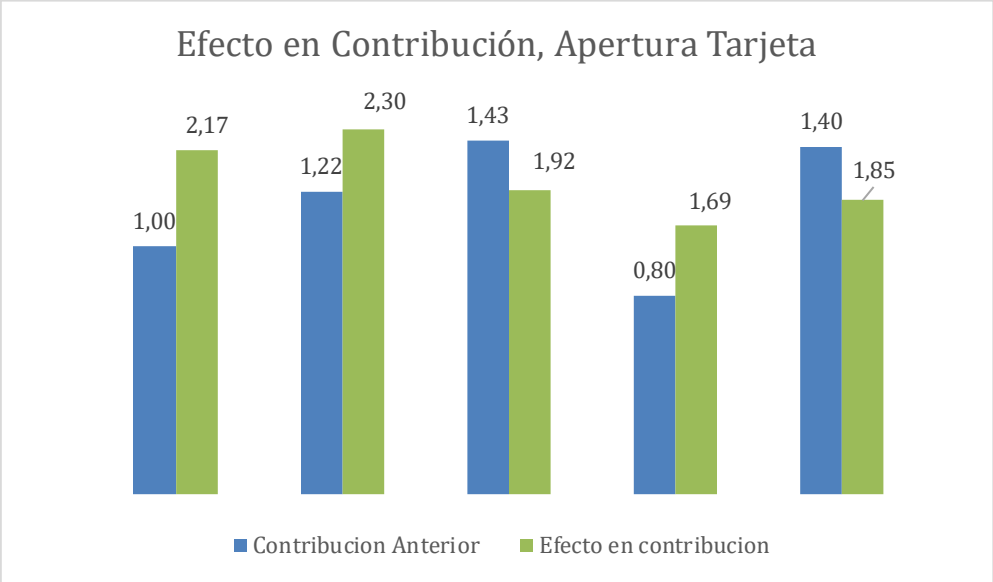


Gráfico 47: Efecto causal para contribución en segmentos, apertura de tarjeta (datos normalizados a la contribución anterior del segmento 1)

Un resultado similar se obtiene para la contribución, no existen grandes variaciones en los efectos causales. Estos varían entre 0.27-0.36 veces la contribución anterior del segmento 1 sin una clara tendencia, a excepción de una leve disminución en los 2 segmentos superiores. Una explicación de esto es que aquellos clientes con mayor propensión aprovechen de mejor forma las ofertas y por eso mismo eran más propensos a obtener la tarjeta. Esto implicaría que de alguna forma las variables utilizadas para los modelos de propensión, alguna de ellas se correlacione con este factor de comportamiento de los clientes, una de las variables que posiblemente lo haga es la de edad, clientes más jóvenes podrían aprovechar mejor las ofertas.

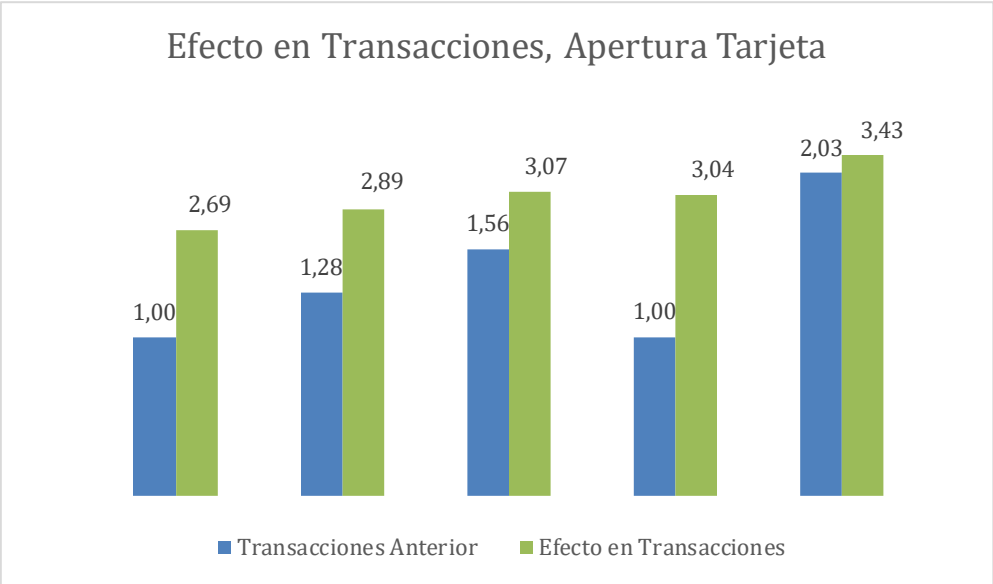


Gráfico 48: Efecto causal para transacciones, apertura de tarjeta (datos normalizados a las transacciones anteriores del segmento 1)

Finalmente se tienen los resultados por transacciones (gráfico 48), se observa una leve alza en el efecto causal en transacciones a medida que se avanza en segmentos. Lo cual al observarlo en conjunto con los resultados anteriores podría ser interesante, pues se ve que los clientes más propensos, al abrir su tarjeta comercial tienen un aumento mayor de transacciones que los otros segmentos, un nivel de gasto similar y una contribución bruta menor.

Los intervalos de confianza se pueden observar en detalle para los efectos promedios y los de segmentos en los anexos 18 y 19, además de los histogramas de los efectos para cada cliente en el anexo 17. Cabe mencionar que todos los resultados, tanto promedios generales como en segmento logran concluir con un 99.9% de confianza que son distintos de 0.

7.2.3 CRUCE DE NEGOCIO: FORMATO 1 A FORMATO 2

El tercer hito que evaluar corresponde al cruce entre negocios, el cual se divide en: cliente que pasa de Formato 1 a Formato 2 y cliente que pasa de Formato 2 a Formato 1. La relevancia de estos hitos recae en que existe la percepción de que el traspaso de clientes entre negocios podría producir canibalización de compras entre estos mismos dado que se comparten categorías de productos, por esto es de interés el efecto causal de estos hitos en los ingresos totales de los 3 negocios agrupados. Otro caso interesante sería ver la existencia de esta canibalización a nivel agregado, lo que se hace en el final de este informe. Los Boxplots de limpieza de datos se encuentran en el anexo 20.

Grupo	Clientes
0	~1,100,000
1	~90,000

Tabla 34: Cantidad de clientes por grupo en, paso de Formato 1 a Formato 2

Para el cálculo de los efectos causales de estos hitos se cuentan con cerca de 90 mil clientes que durante enero del 2017 compraron por primera vez (después de un año) en el negocio de Formato 2 y ya eran clientes de Formato 1 durante el año anterior. Por otro lado, se cuenta con una muestra superior al millón de clientes que no realiza esto. Con estos datos se calculan los modelos de propensión para realizar el matching. Obteniendo los resultados en métricas de balance de la tabla 35.

Modelo	SMD promedio	SMD maximo	SMD std
ANN	0.008	0.022	0.007
Logistica	0.011	0.026	0.009
Forest	0.051	0.118	0.039

Tabla 35: Métricas de balance (SMD) con distintos modelos, Formato 1 a Formato 2

Al igual que en los 2 hitos anteriormente vistos, el modelo de Random Forest es el que genera un peor balance, y nuevamente Red Neuronal y Regresión Logística logran desempeños similares. Esta vez la red neuronal logra un mejor balance en las variables, con menor máximo y menor promedio. Por lo que este modelo es el seleccionado para estimar los efectos causales del paso de Formato 1 a Formato 2. Las variables importantes y su nivel de balance con el método de mejor balance se pueden encontrar en el anexo 21. A continuación, se muestran las métricas de desempeño para clasificación en este hito.

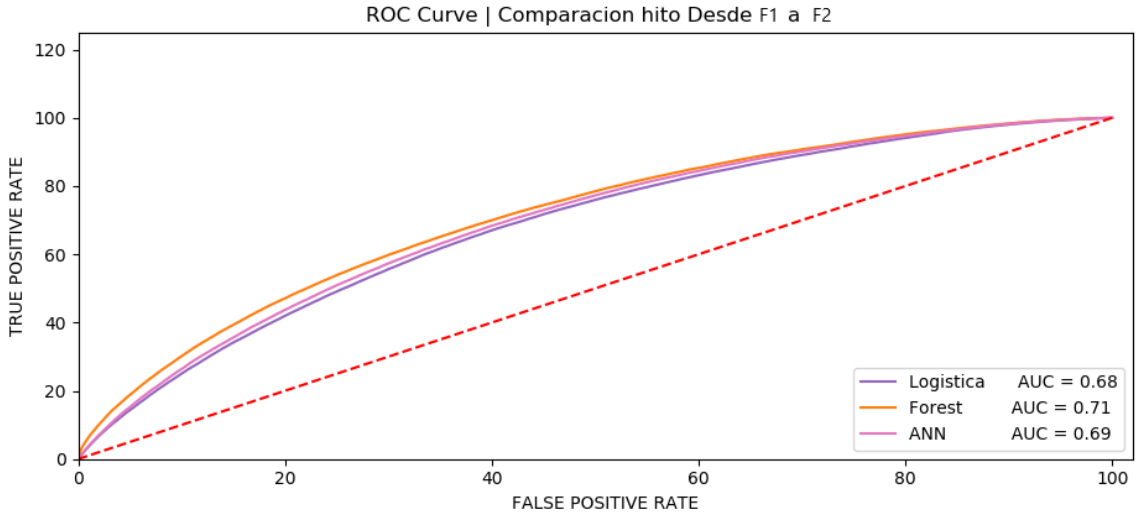


Gráfico 49: Curva ROC, Formato 1 a Formato 2

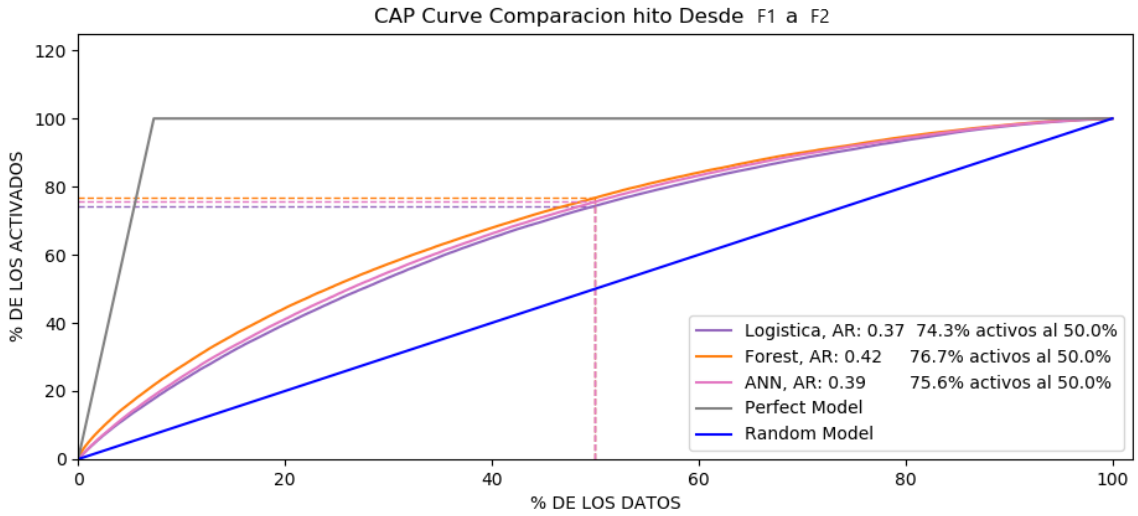


Gráfico 50: curva CAP, hito Formato 1 a Formato 2

Un resultado interesante en este hito es que, si bien nuevamente Random Forest es el que mejores métricas de clasificación obtiene, tanto en la curva como en la curva CAP, esta vez los desempeños son más cercanos. Así se puede ver en el gráfico 49, que la regresión logística solo está 0.01 bajo la red neuronal, la cual a su vez esta 0.02 bajo el Random Forest. Por el lado de la cura CAP, la regresión logística logra identificar al 74.3% de los activados en la mitad de la muestra y la Red Neuronal solo 1.3% menos. Esto resulta interesante pues, estos modelos son más complejos y requieren de más recursos computacionales para su calibración que la Regresión Logística, y aun así logra

resultados similares. De todas formas, el modelo escogido es el de Red Neuronal por obtener las mejores métricas de balance en SMD.

Una vez escogido el modelo se pueden obtener los puntajes de propensión, donde se obtiene la separación de las distribuciones de probabilidad que se muestran en el gráfico 49.

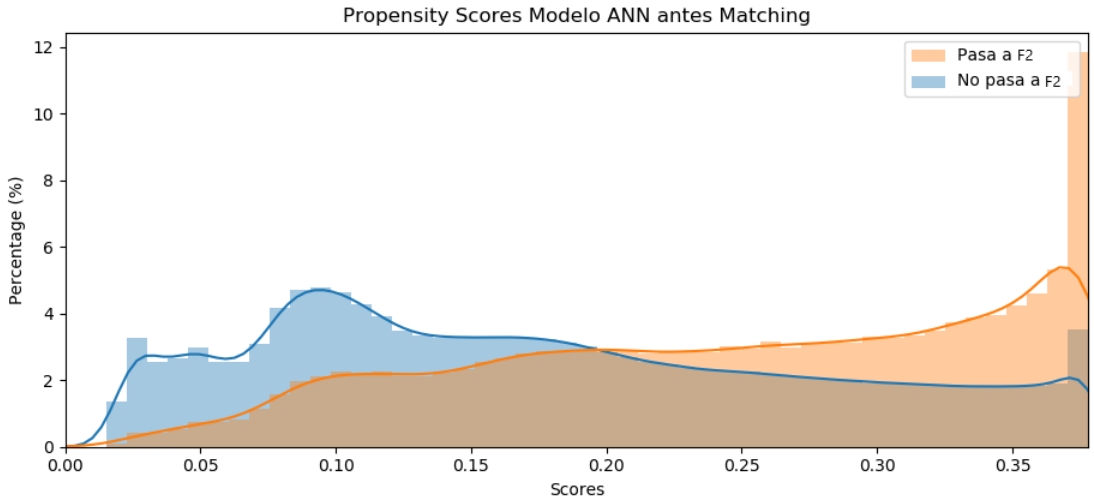


Gráfico 49: Separación de muestras, hito Formato 1 a Formato 2

Como se observa en el gráfico 49 las distribuciones no superan el 40% de probabilidad, quedando tanto los grupos activados como los no activados, bajo el criterio del 50% para predecir a alguien como activado (si es que este fuera un problema de clasificación de clientes). La separación de los otros métodos tanto antes como después del match se encuentra en el anexo 22.

Después de calculados los niveles de propensión se pueden estimar los efectos causales promedios. Que se muestran a continuación en la tabla 36.

Variable	Grupo	Antes	Despues	Diferencia	Diff in diff
Gasto	Activado	1.00	1.80	0.79	0.47
	No activado	1.00	1.32	0.32	
Contribucion	Activado	0.98	1.92	0.94	0.56
	No activado	1.00	1.38	0.38	
Transacciones	Activado	1.01	1.69	0.68	0.40
	No activado	1.00	1.29	0.29	

Tabla 36: Efectos causales de paso de cliente Formato 1 a Formato 2 (normalizados al valor del no activado en la variable)

En la tabla 36 se muestra que ambos grupos parten en un nivel similar de gasto, un nivel similar de contribución y un nivel similar de transacciones. Así por el lado del gasto, los clientes que realizan el hito crecen en 0.79 su gasto versus 0.32 los que no lo hacen. En términos de contribución el diferencial sobre el crecimiento de los no activados es de 0.56. Y Finalmente, aquellos clientes que se activaron tuvieron un efecto de 0.4 en transacciones.

Algo interesante de mencionar acá, es que a pesar de que el efecto en transacciones del hito de pasarse de Formato 1 a Formato 2 es 0.4, el cual es mayor que el efecto de comprar por primera vez online, el hito del canal web tiene un efecto en gasto mayor y un efecto en contribución menor. O sea que el hito de la primera compra web, genera un mayor efecto en gasto que la del cruce de negocio de Formato 1 a Formato 2, con menos transacciones y genera un menor efecto en contribución. Esto se puede deber a que el canal web en general produce menores niveles de margen y además que el negocio de Formato 2 tiene mejores niveles de margen que el de Formato 1, según se muestra en el gráfico 6.

Posterior a la estimación de los efectos causales promedios, se puede obtener los efectos causales en segmentación. El método utilizado para obtener la segmentación es nuevamente el de cortes aleatorios por obtener el mejor nivel de balance, lo que se puede ver en detalle en el anexo 22. Ahí se podrá ver que se obtienen niveles bastante similares de balance con los 3 métodos

Los segmentos quedan constituidos según se muestra en la tabla 37.

Segmento	Cantidad de parejas	Limite izquierdo	Limite derecho	% del total
1	9,536	2%	11%	11%
2	16,415	11%	19%	18%
3	29,553	19%	30%	33%
4	13,047	30%	34%	15%
5	21,259	34%	38%	24%

Tabla 37: Composición de segmentos, caso paso de Formato 1 a Formato 2

Para este caso nuevamente podemos encontrar grandes diferencias en las composiciones de los segmentos. El segmento número 3 es el más grande, en términos del total de la muestra, contando con el 33% del total de los clientes, y el más pequeño es el primero, que tiene solamente un 11% del total de la muestra.

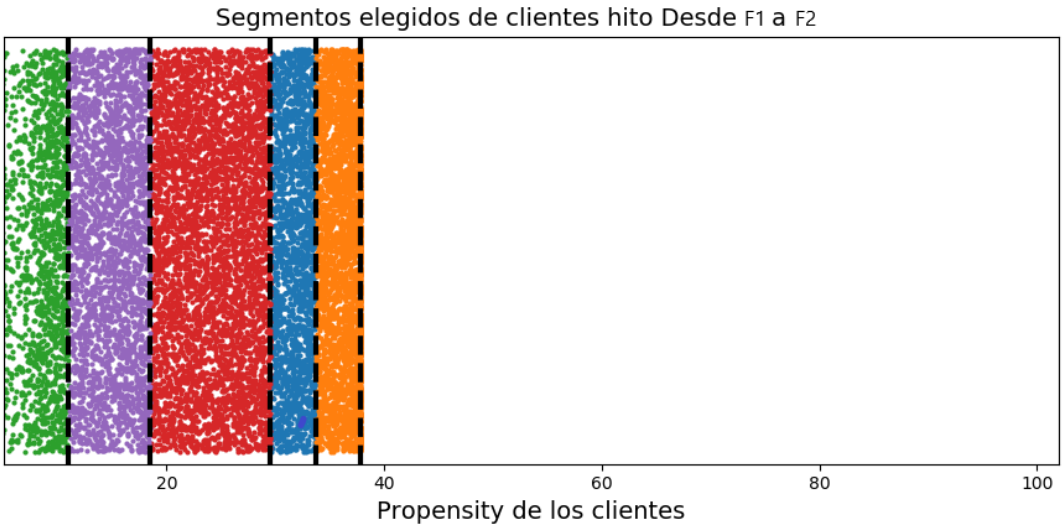


Ilustración 10: Composición de los segmentos de cruce Formato 1 a Formato 2

En la ilustración 10 se puede ver la composición gráfica de los segmentos. Rápidamente resalta al a vista que estos segmentos, a diferencia de los 2 casos anteriores, son más compactos. Esto se puede observar en el último segmento donde los clientes ya no se encuentran dispersos en un gran dominio de probabilidad, si no que la probabilidad más grande no supera el 40%. En la tabla 38 se puede ver cómo se comportan estos segmentos en las variables sociodemográficas.

Variable	1	2	3	4	5
Gasto anterior	1.00	1.92	4.28	7.57	12.38
Edad	1.00	1.20	1.26	1.25	1.24
Hombre	1.00	1.86	2.32	2.29	2.30
Tarjeta	1.00	16.38	80.65	132.27	150.53
Hijos	1.00	1.27	1.36	1.37	1.46
Soltero	1.00	0.85	0.70	0.67	0.59
Zona norte	1.00	1.81	2.00	1.96	2.23
Zona RM	1.00	1.19	1.33	1.44	1.60
Zona Centro sur	1.00	0.81	0.73	0.67	0.57
Zona sur	1.00	0.84	0.71	0.64	0.49
Cliente Formato 3	1.00	2.18	10.18	30.65	60.37

Tabla 38: Promedios en variables según segmento de efecto causal, caso Formato 1 a Formato 2 (Normalizado)

Lo primer que resalta de la tabla 38, es que existen grandes diferencias en niveles de gasto promedio según segmento, el primero segmento parte con 1 de gasto y el último llega hasta los 12.38, lo que deja entrever una fuerte correlación entre nivel de gasto y probabilidad de cruzar de negocio. Por otro lado, y a diferencia de los casos anteriores, los segmentos más altos cuentan con mayor proporción de hombres, lo que se condice

con las concepciones socio-culturales hegemónicas (estereotipos de género). Un resultado también contrario a los anteriores es que la propensión a realizar el hito (y con esto también el segmento) se correlaciona positivamente con la cantidad de hijos y negativamente con ser soltero, lo que tiene sentido pues se trata de un negocio enfocado al Formato 2. Finalmente, teniendo ya los segmentos establecidos, se pueden estimar los efectos causales en segmentación para el gasto, la contribución y las transacciones.

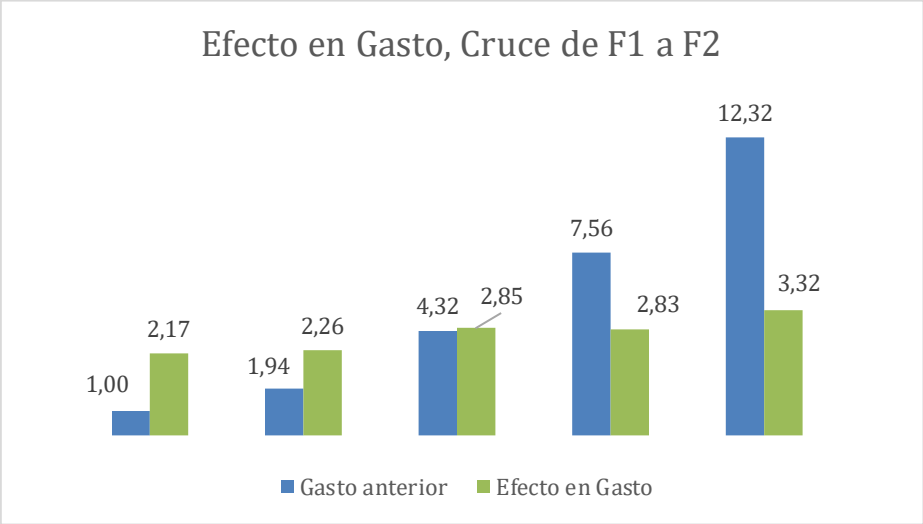


Gráfico 50: Efecto causal en gasto, cruce de Formato 1 a Formato 2 (Normalizado)

En el gráfico 50 se puede observar que a medida que se aumenta de segmento, también se va aumentando de efecto causal en gasto. Para este caso este aumento representa una tendencia marcada, lo que se diferencia de los resultados de los hitos anteriores. Una explicación de esto puede ser las diferencias grandes que existen con respecto al gasto anterior entre los segmentos, el segmento de menor gasto parte en 1 promedio normalizado y el de mayor termina en promedio de 12.32 mil. Estas diferencias pueden también evidenciar diferencias de capacidad de gasto de los clientes. Por lo tanto, los clientes que más gastaban en el holding antes de hacer el cruce también tienden a gastar más en el negocio al que se cruzan.

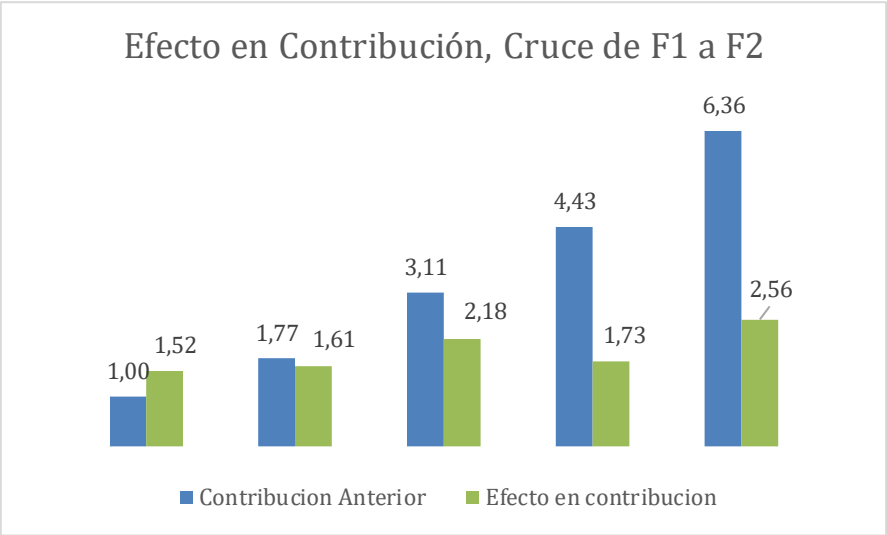


Gráfico 51: Efecto causal en contribución, cruce de Formato 1 a Formato 2 (Normalizado)

Por el lado de la contribución también se observa una tendencia al alza en medida que se sube de segmento, pero esta no siempre se cumple. Esto se puede ver en el caso del segmento 4 que tiene un efecto en contribución menor al segmento 3, pero tiene un efecto en gasto similar. Por lo que obviando lo que sucede en el segmento 4, se ve la tendencia al alza en los resultados.

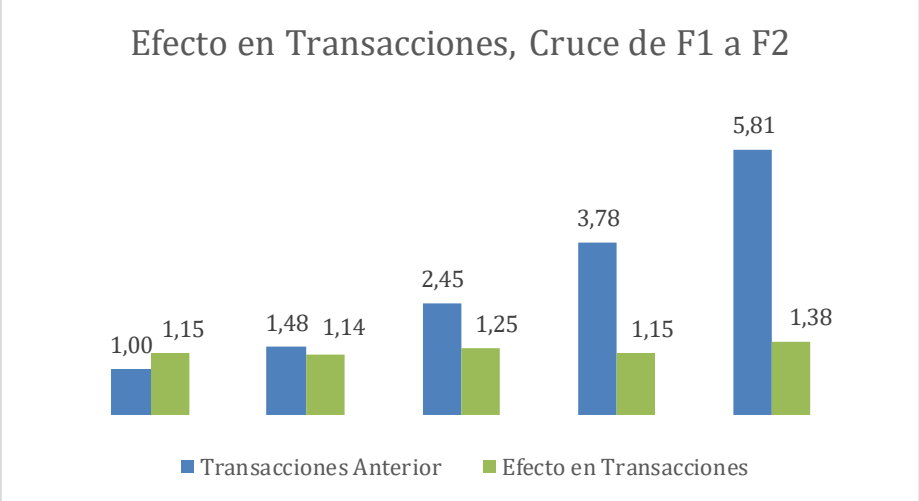


Gráfico 52: Efecto causal en transacciones, cruce de Formato 1 a Formato 2 (Normalizado)

Por el lado de las transacciones, según se muestra en el gráfico 52, se produce un caso similar al de la contribución, pues el segmento 4 tiene un efecto menor al segmento 3, pero el segmento 5 supera el de todos los demás. Si bien los segmentos 1, 2 y 4 tienen un efecto en transacciones similar, el aumento en gasto y en contribución para estos 3 segmentos se mueve en la línea de la tendencia al alza según segmento.

Así es como finalmente para este hito se puede observar claras tendencias al alza en cada efecto causal a medida que se avanza en los segmentos, la explicación más probable de esto es que un cliente que ya gastaba más en los negocios que se encontraba anteriormente, tiene una mayor capacidad de gasto que los clientes que gastaban menos. Además de que a diferencia del hito de la primera compra web, que lo que se cambia es el canal de compra, en este caso también se está cambiando el tipo de negocio agregando el de Formato 2. Si se toma en consideración que las personas tienden a gastar un porcentaje relativamente fijo de sus ingresos en del hogar, un cliente tuviera mayor capacidad de gasto, necesariamente gastaría más en este nuevo negocio que aquellos que tienen menor capacidad a diferencia del hito del canal web donde no necesariamente existe un porcentaje de ingresos que las familias gastan en ese canal.

7.2.4 CRUCE DE NEGOCIO: FORMATO 2 A FORMATO 1

El último hito para estudiar corresponde a la segunda variante de cruce de negocios, aquellos clientes que durante el 2016 fueron clientes del negocio de Formato 2 y no de Formato 1, que se empiezan a comprar en Formato 1 durante enero del 2017.

Grupo	Clientes
0	~400,000
1	~30,000

Tabla 39: Cantidad de clientes en cada grupo, caso Formato 2 a Formato 1

Como se ve en la tabla 39, los clientes que se activan en el negocio de Formato 1 durante enero son cerca de 30 mil personas y se toma una muestra de 400 mil clientes que no se activan. Posteriormente se estiman los modelos de propensión y se obtienen los resultados de balance presentes en la tabla 40.

Modelo	SMD promedio	SMD maximo	SMD std
ANN	0.006	0.021	0.005
Logistica	0.008	0.036	0.009
Forest	0.074	0.210	0.053

Tabla 40: Resultados modelos de propensión en balance SMD, caso Formato 2 a Formato 1

Como se puede observar en la tabla anterior, el modelo de Red Neuronal es el que mejor balance genera, tanto en SMD máximo con promedio. Nuevamente y al igual que en todos los casos anteriores, el balance generado con Random Forest es el de más SMD, siendo su promedio bastante cercano al límite de corte de 0.1, además de tener su máximo por sobre este límite. A continuación, se pueden ver las métricas de clasificación, ROC-AUC y CAP. El balance se calculó sobre las variables más importantes, las cuales se pueden ver en el anexo 28.

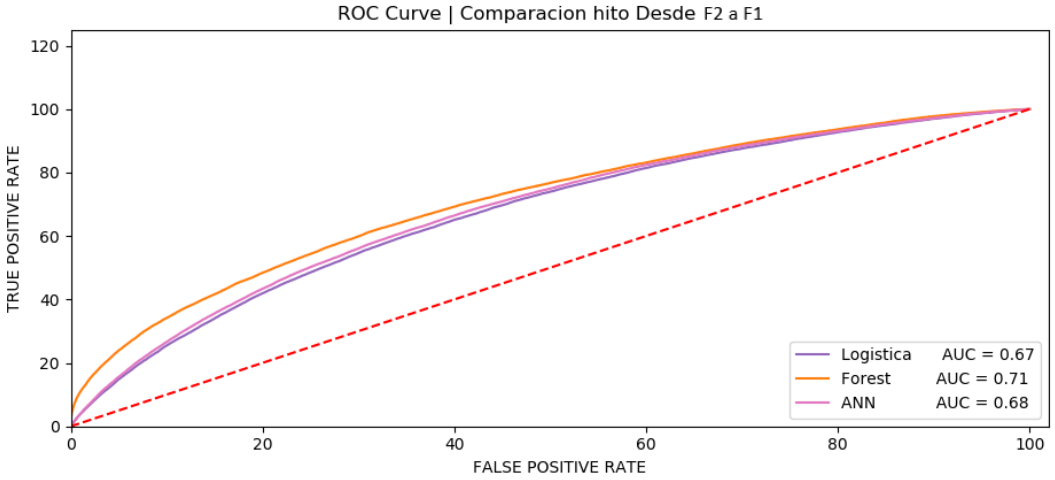


Gráfico 53: Curva ROC-AU, caso cruce de negocio de Formato 2 a Formato 1

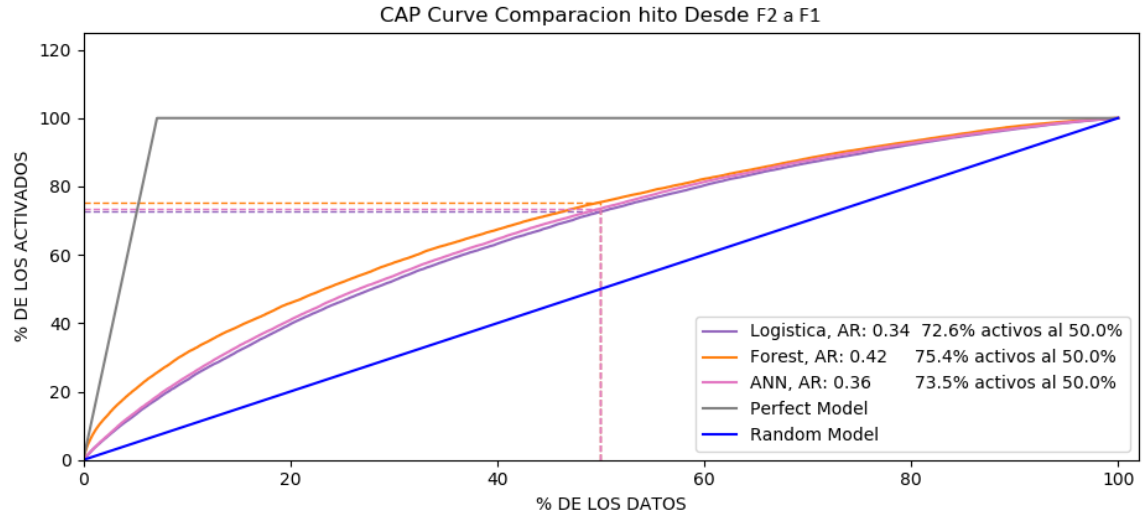


Gráfico 54: Curva CAP, caso cruce de negocio de Formato 2 a Formato 1

Por el lado de las métricas de clasificación, el área bajo la curva del gráfico ROC es mayor, al igual que en todos los casos anteriores, para el modelo de Random Forest. En el segundo lugar se ubica la Red Neuronal y en último la Regresión Logística. Esto se confirma en la curva CAP, donde si bien la proporción de activados identificados al 50% de la muestra es similar para los tres casos, el Random Forest logra identificar de mejor forma a los clientes más propensos a realizar los hitos.

Finalmente, el modelo escogido es el de Red Neuronal. Con las propensiones generadas por este modelo se pueden ordenar las distribuciones de probabilidad de cada grupo según se muestra en el gráfico 55.

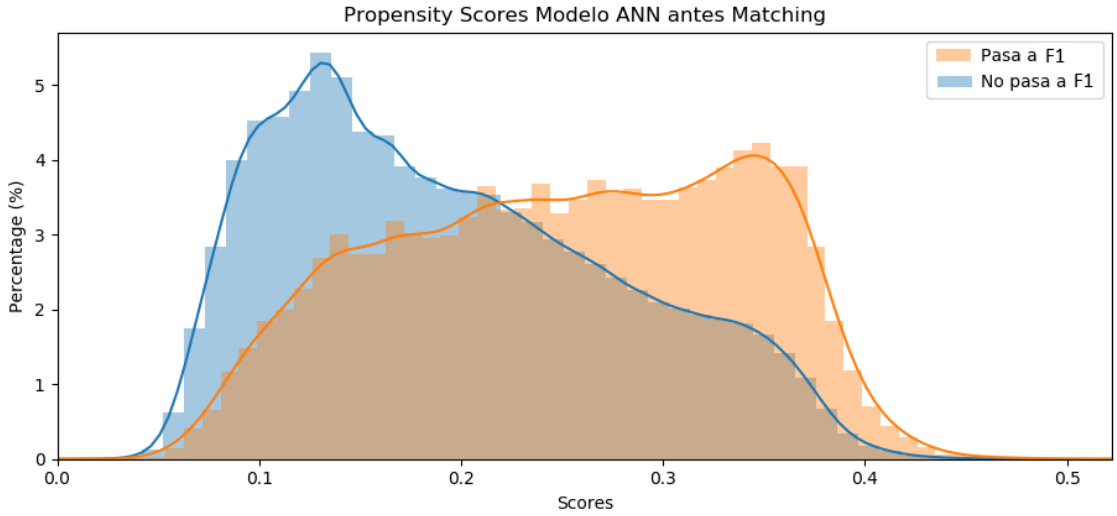


Gráfico 55: Separación de las distribuciones de probabilidad, caso cruce de negocio de Formato 2 a Formato 1

Una vez generados los scores de propensión se puede proceder a realizar el match, donde a cada cliente se le busca su no activado de menor distancia de propensión, siempre y cuando no supere el umbral del caliper. Cabe recordar que este match se realiza con 1 no activado para cada activado y sin reemplazo.

Se obtienen los resultados de efectos causales que se muestran en la tabla 41, se cuenta con los resultados en gasto, contribución y transacciones.

Variable	Grupo	Antes	Despues	Diferencia	Diff in diff
Gasto	Activado	1.00	1.93	0.93	0.43
	No activado	1.00	1.50	0.50	
Contribucion	Activado	1.00	1.72	0.72	0.42
	No activado	1.00	1.29	0.29	
Transacciones	Activado	0.99	1.99	1.00	0.41
	No activado	1.00	1.59	0.59	

Tabla 41: Resultados de efectos causales, hito de cruce de negocio de Formato 2 a Formato 1. (Normalizado al valor de la variable en el caso no activado)

Como se muestra en la tabla, el efecto causal de este hito en gasto es inferior a los 0.43 veces el promedio de los no activados, lo que representa el hito con menor efecto causal de los estudiados. Ambos grupos parten desde un mismo nivel de gasto. Por el lado del efecto causal en contribución, vemos que el aumento es de 0.42, similar al efecto en gasto y transacciones (en valores de los promedios respectivos).

Por el lado de las transacciones, se ve que también en línea con los demás efectos, 0.41 transacciones promedios en comparación con el grupo de no activados matcheados. Esto representa el menor efecto causal visto en transacciones.

Una vez estimado los efectos causales en promedio, se puede proceder a la estimación de los segmentos para poder medir efectos heterogéneos entre grupos de clientes. El método con el cual se generan los segmentos es el de cortes aleatorios pues es el que genera mejor balance.

Segmento	Cantidad de parejas	Limite izquierdo	Limite derecho	% del total
1	7,936	4%	18%	24%
2	7,289	18%	25%	22%
3	5,687	25%	30%	18%
4	8,035	30%	36%	25%
5	3,541	36%	49%	11%

Tabla 42: Composición de los segmentos, caso cruce de negocio de Formato 2 a Formato 1.

Como se muestra en la tabla 42, los segmentos nuevamente no quedan muy distintos del caso de segmentos de quintiles. Solamente existe un segmento que tiene considerablemente menos datos que es el número 5, con solamente el 11% de la muestra. El valor máximo de propensión obtenido es el de 49%, por lo que todos los clientes quedan por debajo del límite del 0.5. Gráficamente se puede ver cómo quedan ordenados los segmentos en la ilustración 11.

Ilustración 11: Composición gráfico de los segmentos

Para poder entender mejor la composición y características sociodemográficas de los clientes que componen cada segmento, se presentan los valores promedios de algunas variables en la tabla 43.

Variable	1	2	3	4	5
Gasto anterior	1.00	1.41	1.86	2.38	2.97
Edad	1.00	0.99	0.97	0.95	0.87
Hombre	1.00	0.58	0.62	0.59	0.54
Tarjeta	1.00	3.33	4.97	6.63	7.96
Hijos	1.00	0.94	0.88	0.84	0.65
Soltero	1.00	1.06	1.09	1.09	1.23
Zona norte	1.00	0.69	0.48	0.39	0.25
Zona RM	1.00	1.14	1.57	1.75	2.22
Zona Centro sur	1.00	0.96	0.71	0.64	0.49
Zona sur	1.00	1.44	1.30	0.96	0.63
Cliente SM	1.00	6.26	20.10	39.91	44.94

Tabla 43: Promedios en variables según segmento de efecto causal, caso hito Formato 2 a Formato 1 (normalizado)

En la tabla 43 se puede confirmar que al igual que en todos los hitos anteriores, a medida que se avanza de segmento y a su vez se avanza de propensión, también se va aumentando la proporción de clientes de ese grupo que pertenecen a la región metropolitana y disminuyen todas las demás. El promedio de edad es más alto que para el resto de los hitos, pero esto se puede deber a que justamente los clientes de este negocio tienden a tener más edad promedio. Sin embargo, los clientes más jóvenes dentro de este grupo de mayor edad, tienden a ser más propensos a cruzar al negocio de Formato 1. La proporción de clientes con tarjeta aumenta significativamente con los segmentos, lo que resulta similar a los otros hitos. Otro punto interesante es que si bien el segmento 1 (el de menor propensión).

En general las tendencias, si bien son menos marcadas, presentan el sentido contrario al caso de cruce de negocio de Formato 1 a Formato 2. Donde eran clientes de mayor edad, con más hijos y menor proporción de solteros los más probables a llegar al negocio de Formato 2, y aquí se produce lo contrario, los más jóvenes y con menos hijos son los más probables a ir al negocio de Formato 1.

Así es como, ya armados los segmentos, se puede pasar a la estimación de los efectos causales de cada segmento.

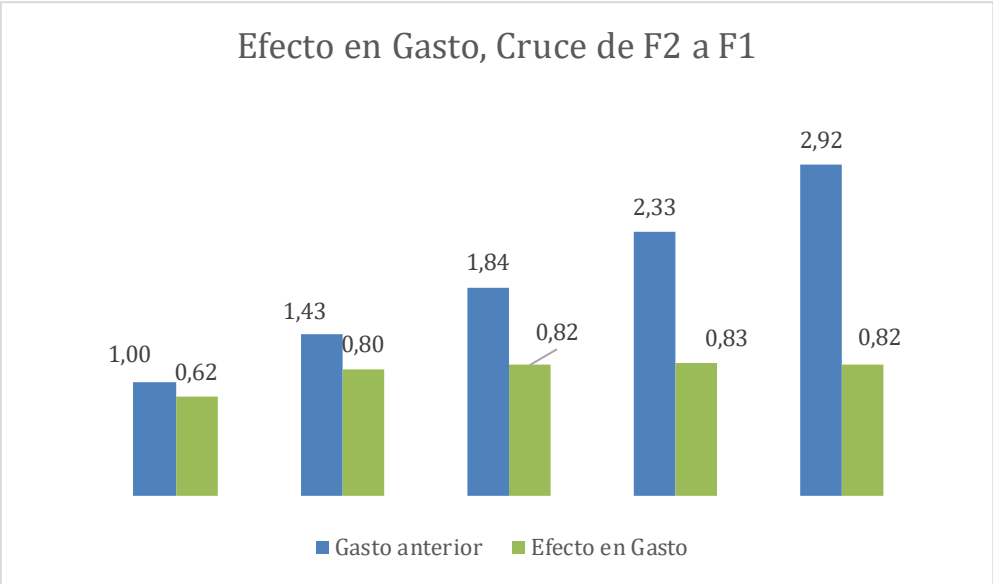


Gráfico 56: Efecto causal en gasto, cruce de negocio de Formato 2 a Formato 1 (Normalizado)

En el gráfico 56 se puede ver que no hay una clara relación entre gasto anterior y efecto causal, de hecho, desde el segmento 2 al 5, el efecto causal varia de 0.80 a 0.82. Esto marca una diferencia con respecto del hito contrario, el cruce de Formato 1 a Formato 2, allí se presenciaba una tendencia de los clientes que cruzaban de negocio. Esto parece indicar que, sin importar la capacidad de gasto de los clientes, el efecto de los clientes que se activan es relativamente constante.

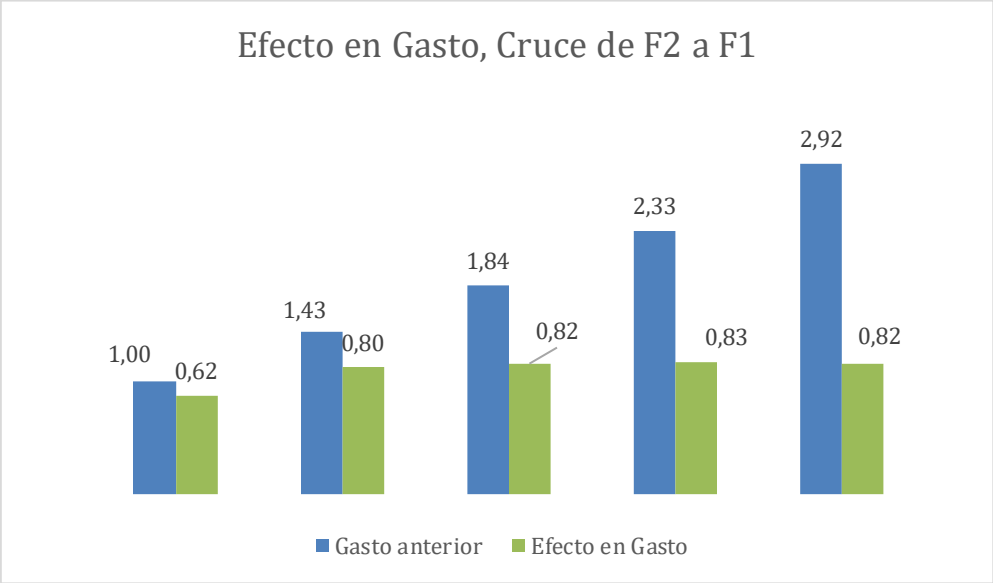


Gráfico 57: Efecto causal en contribución, cruce de negocio de Formato 2 a Formato 1 (Normalizado)

Por el lado de los resultados en contribución, no hace más que confirmar lo que ya se vio en los resultados del gasto. Los clientes con mayor gasto anterior tampoco

necesariamente tienen un efecto causal mayor en contribución. Esto es esperable sabiendo que el efecto en gasto también se mantiene relativamente constante.

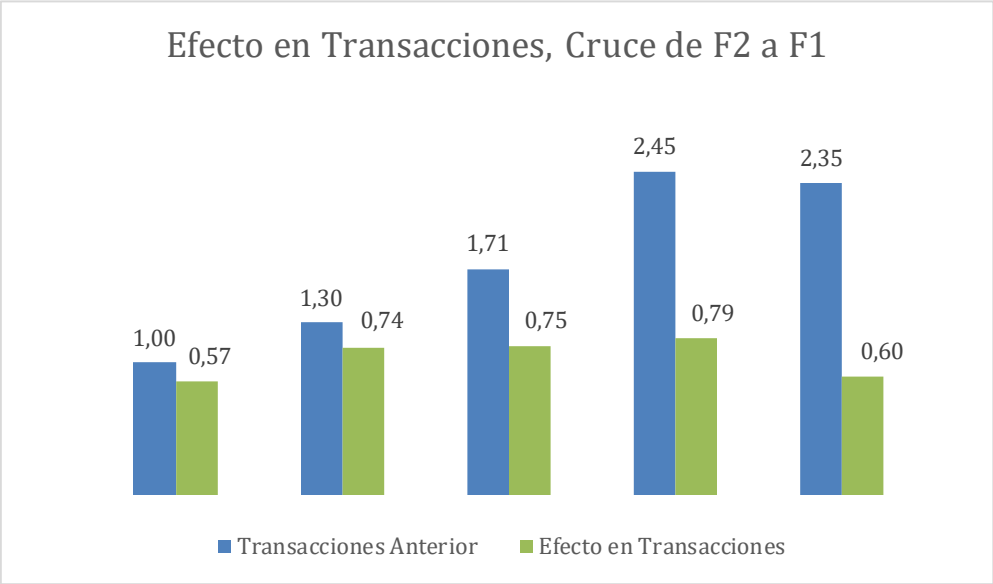


Gráfico 58: Efecto causal en transacciones, cruce de negocio de Formato 2 a Formato 1 (Normalizado)

Sin embargo, en el efecto causal de transacciones se puede tener un resultado un poco más interesante, pues a pesar de que el efecto causal en transacciones de los segmentos 2, 3 y 4 son similares, los segmentos 1 y 5 tienen 0.2 promedios de transacciones menos en promedio que los otros. Si se toma en consideración que el efecto causal en gasto para todos los segmentos es similar, entonces necesariamente el ticket promedio de las transacciones nuevas es más alto para los clientes del segmento mayor, esto porque gastan lo mismo que los demás, pero en menos transacciones.

Finalmente, de este hito es el que se puede obtener menos tendencias claras a la hora de ver como varían los resultados en los distintos segmentos. Para ver los resultados de significancia, intervalos de confianza e histogramas de efectos causales por cliente, se pueden ver los anexos 32 y 33.

7.3 PROPENSIÓN PARA RECOMENDACIONES

Una parte importante del esquema de gestión propuesto en esta memoria es la propensión de los clientes a realizar los hitos en la actualidad. En las secciones anteriores se estimaron estos puntajes de propensión, pero con la finalidad de utilizarlos como parte de la metodología de Propensity Score Matching. En esta sección estos mismos modelos se calibran con datos actuales para estimar las propensiones de los clientes con un fin de gestión, o sea discriminar niveles de inversión y priorización de clientes para la activación de estos.

Así es como se calibran los modelos con los datos actualizados de la siguiente forma:

- La ventana de observación del hito será del mes de enero 2019
- La ventana de las variables será un año anterior, desde enero 2018 a diciembre 2018.

La razón de recalibrar los modelos con datos y observaciones a diciembre del 2018 es que en caso de que hayan cambiado las correlaciones entre las variables, esto se pueda tomar en consideración implícitamente en los modelos. Dado que existen 2 años de diferencia en los periodos de las estimaciones de los modelos utilizados para los efectos causales y los de gestión, el supuesto de que estas relaciones puedan haber cambiado en ese tiempo es posible. Por otro lado, en caso de que no hayan cambiado, los modelos calibrados en enero del 2019 no deberían ser distintos a los de enero del 2017, por lo que no se pone en riesgo ningún resultado de la memoria.

Para estimar los modelos de propensión calibrados en enero 2019, se separa la muestra de datos de clientes que durante ese mes se activa o no se activa en el hito específico. Las separaciones serán: una muestra del 80% en el cual se realiza el CrossValidation simultáneamente con el GridSearch con la finalidad de quedarse con el modelo que mejor aprenda de los datos en métrica de ROC-AUC y otro 20% se utiliza para confirmar que el modelo en específico no esté sobre ajustado.

En esta parte resulta fundamental asegurarse que los modelos no estén sobre ajustados, dado que se utilizan como métrica para priorizar recomendaciones de gestión de los hitos. De todas formas, este ejercicio no se realiza con la finalidad de discriminar modelos, pues el modelo a utilizar será el mismo que se utiliza para estimar los puntajes de propensión de los modelos en cada caso.

Los gráficos donde se muestran las curvas ROC-AUC y CAP de cada caso se encuentran en anexos. En las métricas de desempeño encontradas en anexos se puede observar que en casi todos los hitos los en el periodo de testeo, todos los modelos tienen desempeños muy similares (Anexos desde el 34 al 41). De todas formas, el seleccionado es el modelo con el que se estimó el efecto causal en cada hito.

7.4 ANALISIS DE RENTABILIDAD

Para el análisis de rentabilidad potencial de la gestión de los hitos se observa la distribución de probabilidades de realización de los hitos en la actualidad.

7.4.1 CLIENTES POTENCIALES

Los clientes potenciales de cada hito corresponden a aquellos clientes que hayan comprado al menos una vez en alguno de los 3 negocios en el periodo julio 2018-junio 2019.

Hito	Clientes potenciales
Compra web	~4,000,000
tarjeta	~3,000,000
Cruce F1 a F2	~1,000,000
Cruce F2 a F1	~1,500,000

Tabla 44: Clientes que aún no realizan los hitos a junio 2019.

Como se observa en la tabla 44, el hito de primera compra web es que tiene la mayor cantidad de clientes potenciales, llegando a superar los 4 millones. Por otro lado, los hitos de cruce de negocio son los que tienen menor cantidad, pues cumplen la condición más difícil que es ser cliente de un negocio y no del otro.

A continuación, se muestra la distribución de probabilidad de realizar la primera compra web en julio 2019, las líneas azules representan los límites de los segmentos (los otros hitos están anexo 42).

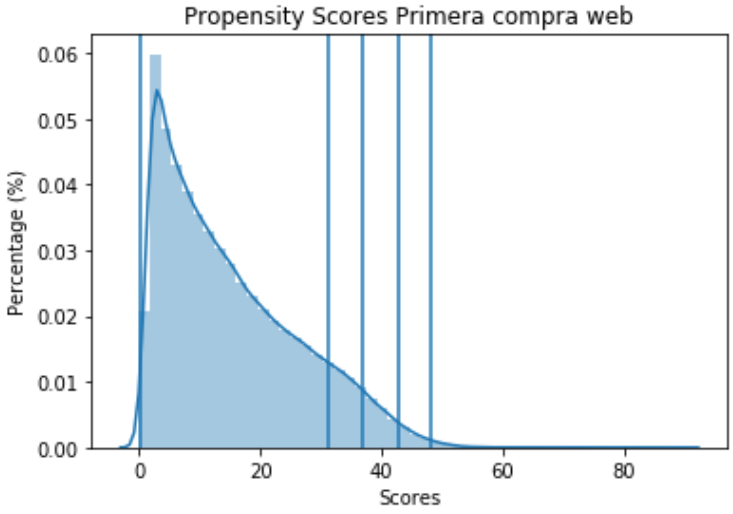


Gráfico 59: Distribución de probabilidad de realizar la primera compra web en julio 2019.

Como se puede observar el segmento más grande en cantidad de cliente es el primero, y una proporción bastante baja de clientes logra pertenecer al último. Esto se puede confirmar en la tabla 45.

Primera compra web				
Segmento	Clientes	Propension	Propension ponderada	Incremento Contribución por cliente Normalizada
1	~4000000	12.44%	0.87%	29.40%
2	~300000	33.89%	2.37%	23.72%
3	~200000	39.40%	2.76%	19.09%
4	~50000	44.99%	3.15%	16.72%
5	~16000	50.93%	3.56%	14.70%

Tabla 45: Resumen de segmentos de primera compra web en julio 2019. (Normalizado a la contribución promedio de los clientes con tarjeta utilizados en el match)

En la tabla 45 se aprecia que, efectivamente el segmento más grande es el 1, que agrupa a más de 4 millones de clientes, con una propensión promedio de 12.44%. En la tabla se muestra la propensión ponderada, que es la propensión tal que hace que la gente que se active de forma natural en julio del 2019 sea similar a la gente que se activa en enero del 2019. Esta corrección se hace porque en general estos son hitos poco probables, como se vio en la tabla 4, solamente un 1% de los clientes que podría activarse en el hito lo hace. Por lo tanto, de suceder realmente las propensiones calculadas en los modelos, más de 700mil clientes se activarían en la primera compra web, cuando realmente esto lo hace solamente del orden de las 50 mil personas. Esta diferencia se produce por el rebalanceo de clases con el cual se calibraron los modelos, como se explica en el desarrollo de la metodología se usó un ratio de 1:4 porque disminuía el desbalance de variables, pero la proporción real es 1:100.

Considerando la propensión ponderada, entonces en general la probabilidad más que una herramienta para diferenciar rentabilidades sirve para priorizar clientes, gestionando a los más probables. Finalmente se puede observar que de activarse los clientes en este hito se podría incrementar en cerca de un 23% la contribución de cada cliente que se active, a lo que habría que restarle el costo de la gestión que podría venir en la forma de un descuento. Uno de los supuestos sumamente fuertes en esta memoria es que esta evaluación de efecto causal sobre contribución y otras variables se podría repetir en el futuro en caso de que los clientes sean gestionados para hacer los hitos.

Hito	Clientes	Propensión ponderada	Incremento Contribución por cliente Normalizada
Primera compra web	~4,000,000	2.50%	23%
Apertura de tarjeta	~3,000,000	0.70%	169%
Cruce de Negocio F1 a F2	~1,000,000	6.20%	56%
Cruce de Negocio F2 a F1	~1,500,000	3.60%	42%

Tabla 45.2: Resumen de rentabilidades potenciales por hito (Normalizada)

Con respecto a los otros hitos. La apertura de tarjeta genera, al igual que el hito de primera compra web, uno de los mayores aportes en rentabilidad, un aumento del 169% de su contribución anterior. Para considerar un análisis más cercano a la realidad hay que tomar en consideración los costos derivados de la gestión de los hitos (los descuentos o cupones entregados por email marketing y otros costos asociados)

Finalmente, los hitos de cruce de negocio son los que generan menor potencial incremental en contribución, y esto es justamente porque tienen una cantidad de clientes mucho menor que el de los hitos anteriores. Ambos incrementales potenciales rondan los 30 mil millones de pesos.

7.5 ANÁLISIS DE ROBUSTEZ

Basar un modelo de gestión en priorización de activación de hitos en la rentabilidad esperada de los mismos requiere que estas estimaciones de efecto causal sean lo más real posibles. Por esto mismo es que es importante realizar un análisis de robustez de los resultados. Por lo tanto, se observa como varían las estimaciones al cambiar ciertos parámetros de la metodología.

Algunas de las cosas que se pueden cambiar para iterar la metodología son:

- Número de no activados que se empareja a cada activado
- Matching con reemplazo o sin reemplazo
- Modelo de estimación de puntajes de propensión
- Ventana de estimación de los efectos

A priori se considera que lo que más podría cambiar la estimación de los efectos causales es la ventana de estimación de los efectos, por ejemplo, envés de ver a los clientes que se activan en enero del 2017 se podría medir con aquellos que se activan en enero del 2018 u otro mes. Otro aspecto que podría cambiar los resultados considerablemente es el número de clones utilizados. Es importante obtener un resultado consistente al aumentar el número de clones, pues a medida que se aumenta esto, se obtienen resultados menos variantes al momento de cambiar la muestra de clientes, pero serán resultados levemente más segados.

7.5.1 NÚMERO DE NO ACTIVADOS

Con la finalidad de poner a prueba la robustez de los resultados es que se iterará la metodología de la memoria, variando el número de no activados emparejado a cada activado. Se muestra solamente los resultados de efectos promedio, pues el resultado en segmentos varía en cada iteración, dado que los segmentos son distintos.

Se prueban 3 casos nuevos, la iteración de la metodología con 2, 3 y 4 no activados por activado y también se muestra el resultado que ya se tenía.

7.5.1.1 CALIDAD DEL EMPAREJAMIENTO

Primero se muestra la calidad del emparejamiento logrado en cada iteración como el promedio de la diferencia de medias estandarizadas de cada variable entre los 2 grupos: activados y no activados.

	SMD promedio				
Hito	1	2	3	4	Modelo
Primera compra web	0.011	0.009	0.010	0.011	Red Neuronal
Apertura de tarjeta	0.009	0.007	0.007	0.006	Regresion logística
Cruce F1 a F2	0.008	0.011	0.011	0.012	Red Neuronal
Cruce F2 a F1	0.006	0.008	0.008	0.009	Red Neuronal

Tabla 46: SMD promedio en cada iteración.

Se iteró el número de no activados por activado utilizando los puntajes de propensión entregados por el mejor modelo en la iteración de los resultados principales de la memoria. Así, en 3 de los 4 hitos se utiliza Red Neuronal y en 1 la regresión logística.

En la tabla 46 se puede observar que en todos los casos de la iteración se logra mantener los promedios de SMD bajo el punto de corte de 0.1. Por otro lado, se puede observar que a medida que se aumenta el número de no activados no se empeora este indicador necesariamente por lo que el ejercicio de esta memoria utilizando cualquiera de estas proporciones de clases de cliente habría sido válido desde el punto de vista de balance.

7.5.1.2 ROBUSTEZ DEL EFECTO EN GASTO

En esta ocasión corresponde hacer el mismo análisis anterior, pero observando los efectos promedios en gasto envés de la diferencia de SMD.

	Gasto				
Hito	1	2	3	4	Modelo
Primera compra web	0.32	0.32	0.32	0.32	Regresion logística
Apertura de tarjeta	2.65	2.65	2.68	2.67	Red Neuronal
Cruce TD a MH	0.47	0.47	0.47	0.48	Regresion logística
Cruce MH a TD	0.48	0.47	0.47	0.48	Regresion logística

Tabla 47: Efecto causal promedio en gasto en cada iteración (Normalizado al gasto anterior)

Como se puede comprobar en la tabla 47, los resultados no tienen grandes variaciones al utilizar cada vez más clientes no activados para las comparaciones. En todos los hitos sucede un fenómeno similar, no existe ninguna tendencia clara a aumentar o disminuir el efecto causal, todo lo contrario, algunas veces sube o baja del punto base del matching realizado con 1 solo cliente no activado por cada activado.

7.5.1.3 ROBUSTEZ DEL EFECTO EN CONTRIBUCIÓN

Por el lado de la contribución se hace el mismo análisis que el caso anterior, pero comparando el margen bruto.

	Contribucion				
Hito	1	2	3	4	Modelo
Primera compra web	0.25	0.24	0.24	0.24	Regresion logística
Apertura de tarjeta	1.68	1.70	1.72	1.72	Red Neuronal
Cruce TD a MH	0.57	0.59	0.59	0.59	Regresion logística
Cruce MH a TD	0.47	0.44	0.45	0.46	Regresion logística

Tabla 48: Efecto causal promedio en contribución en cada iteración (Normalizado a la contribución anterior)

En la tabla 48 se puede apreciar que, al igual que en el caso del gasto, al aumentar el número de no activados los resultados no tienen grandes desviaciones en la contribución. Las variaciones se mantienen dentro del margen del intervalo de confianza del resultado promedio base.

7.5.1.4 ROBUSTEZ DEL EFECTO EN TRANSACCIONES

Finalmente, el último efecto causal a revisar es el de las transacciones.

	Transacciones				
Hito	1	2	3	4	Modelo
Primera compra web	0.21	0.21	0.21	0.21	Regresion logística
Apertura de tarjeta	2.20	2.18	2.18	2.16	Red Neuronal
Cruce TD a MH	0.40	0.41	0.41	0.41	Regresion logística
Cruce MH a TD	0.42	0.42	0.42	0.43	Regresion logística

Tabla 49: Efecto causal promedio en transacciones en cada iteración (Normalizado a la transacción anterior)

Nuevamente se puede confirmar lo que se ha visto en todos los casos anteriores, los resultados en transacciones no varían más allá del segundo decimal con respecto del caso basal. Lo que habla bien a cerca de la robustez de la evaluación de los efectos causales y se mantienen dentro de los intervalos de confianza del caso basal estudiando en los resultados de cada efecto.

Tomando en consideración que todos los resultados se mantienen dentro de márgenes razonablemente pequeños del caso original, es que se puede concluir que los resultados son robustos por el lado del número de clientes no activados utilizados en los emparejamientos.

8.5.2 DISTINTOS MODELOS

Por otro lado, otro posible factor a cambiar es el modelo utilizado para la estimación de los efectos causales. En la memoria se utilizó el modelo que generará el mejor balance post match para cada hito, por lo que ahora se prueba iterar todos los resultados utilizando el segundo mejor modelo en balance para cada hito, pero no se varía el número de clientes tratados por cada tratado, pues ese ese el análisis anterior.

	Transacciones	
Hito	SMD base	SMD otro modelo
Primera compra web	0.011	0.017
Apertura de tarjeta	0.009	0.018
Cruce F1 a F2	0.008	0.012
Cruce F2 a F1	0.006	0.010

Tabla 50: SMD promedio según variación de modelo utilizado.

Las diferencias de medias estandarizada promedio para cada hito variando los modelos utilizados están en la tabla 48. Ahí se puede observar que en general utilizando el segundo mejor modelo en SMD, se tienden a obtener una peor diferencia de medias, pero aún se mantiene bajo el nivel de rechazo de 0.1.

Los modelos utilizados son los segundos mejores para cada hito. Así, para primera compra web, cruce de Formato 1 a Formato 2 y cruce de Formato 2 a Formato 1 se utiliza una regresión logística y para la apertura de tarjeta se usa una Red neuronal.

Hito	Promedio	Lim izquierdo	Lim derecho	Otro modelo
Gasto				
Primera compra web	0.92	0.83	1.00	1.06
Apertura de tarjeta	0.84	0.81	0.87	0.82
Cruce F1 a F2	0.47	0.45	0.49	0.47
Cruce F2 a F1	0.43	0.38	0.49	0.48
Contribucion				
Primera compra web	0.23	0.19	0.26	0.25
Apertura de tarjeta	1.69	1.58	1.81	1.68
Cruce F1 a F2	0.56	0.53	0.60	0.57
Cruce F2 a F1	0.42	0.35	0.49	0.47
Transacciones				
Primera compra web	0.51	0.45	0.57	0.58
Apertura de tarjeta	0.82	0.79	0.85	0.82
Cruce F1 a F2	0.40	0.38	0.42	0.40
Cruce F2 a F1	0.41	0.36	0.46	0.42

Tabla 51: Robustez según segundo mejor modelo para cada hito (Normalizado a la variable pre hito)

En la tabla 51 se muestran los resultados promedios iniciales (columna promedio) que corresponde al efecto causal con 1 cliente no tratado por cada tratado y el mejor modelo en SMD (caso original), también se muestra los límites de los intervalos de confianza para ver si es que el promedio obtenido por el otro modelo se encuentra fuera o dentro del intervalo de confianza anterior. Como se puede observar en la tabla, en general los resultados cambiando el modelo utilizado, no tienen grandes variaciones, todos se mantienen positivos y en el mismo orden de magnitud. Por otro lado, en prácticamente todos los casos el resultado promedio nuevo está dentro del intervalo de confianza del resultado basal, excepto en el caso de la primera compra web en que el nuevo resultado es mayor que el primero obtenido. No existe una tendencia clara de si al cambiar el modelo por uno distinto se tiende a aumentar o disminuir el efecto causal estimado.

Finalmente, si bien algunos resultados se mueven hasta fuera de los intervalos de confianza al cambiar el mejor modelo, no se produce un cambio fundamental del efecto causal como lo fuera un cambio de signo. Por lo tanto, considerando el análisis de la variación del número de emparejados por cada activado y el cambio del modelo con el cual se estiman los efectos causales, se puede concluir que los resultados son robustos en ambas variantes.

8. RECAPITULACIÓN METODOLÓGICA

Para poder estimar los efectos causales de los hitos corporativos en ausencia de experimentación, se utilizó la metodología de Propensity Score Matching (PSM), que mediante modelos de propensión busca hacerse cargo de problema de endogeneidad a la hora de estimar los efectos dado la autoselección de los clientes a realizar los hitos. Con la metodología se logra encontrar submuestras de clientes que tienen las mismas variables observables, logrando así controlar el efecto de autoselección.
Los primeros pasos metodológicos son los siguientes:

1) Primero estimar modelos de propensión de la realización de los hitos con clientes que se hayan activado en el mes de enero del 2017.

2) Una vez estimado los modelos, se calculan los puntajes de propensión a realizar los hitos de todos los clientes (los que se activaron y no activaron).

3) Después se procede a encontrar clientes (1,2 o 3) que estén dentro de la vecindad de probabilidad de los clientes activados. Se itera por cada cliente activado.

4) Se obtiene una muestra de clientes comparables o "clones" de los activados, que en promedio tienen las mismas distribuciones de variables observables

5) Para decidir las variables más importantes de los clientes se utiliza un Random Forest que entrega la importancia relativa entre variables y un modelo Logit con la importancia como el valor absoluto del coeficiente relacionado a la variable.

6) La calidad de emparejamiento se mide con la métrica de Diferencia de Medias Estandarizadas (SMD), si es que para las variables más importantes este valor es menor a 0.1, entonces se considera OK el emparejamiento.

7) Los puntos 1, 2, 3, 4 y 6 se iteran para cada hito y se prueban 3 modelos de propensión distintos (Random Forest, Regresión Logística y Red Neuronal). Seleccionando aquel modelo con menor SMD promedio de las variables más importantes.

8) Los resultados se calculan como diferencia en diferencia de los activados contra los no activados similares.

9) Para los efectos heterogéneos se prueban 3 métodos de segmentación de clientes sobre con respecto a su puntaje de propensión. Se selecciona el de menor SMD promedio en los segmentos.

10) Finalmente se calcula el efecto causal con diferencia en diferencia pero usando los activados y no activados que caen dentro de cada segmento.

9. CONCLUSIONES

9.1 Sobre los resultados

Por el lado del negocio, una de las conclusiones relevantes es que todos los hitos estudiados en la memoria resultan en un efecto positivo en el gasto, contribución y transacciones de los clientes que se activaron. El que genera un efecto mayor en gasto es el hito apertura de tarjeta, donde en casi todos sus segmentos se obtiene efectos de aumento de más de un 260% de gasto. En segundo lugar y tercer lugar empatados están los hitos de paso de F1 a F2 y de F2 a F1 con 47% en ambos casos. Finalmente, el menor efecto causal corresponde a la primera compra web, logrando un incremento de 32% pesos con respecto al incremento percibido por los clientes comparables.

Por el lado de los resultados en contribución, los efectos cambian de orden, pues si bien la apertura de tarjeta implica un aumento en el margen de 170%, el segundo y tercer lugar no están empatados en contribución, pues de F1 a F2 implica un aumento de contribución de 58%, el caso contrario solo es de 47%. Esto se podría deber a que el canal de venta F2 es un canal en general más rentable en margen bruto que el de F1. Finalmente se tiene en el último lugar a la primera compra web con un 24%

Sobre las segmentaciones de los clientes, cada grupo logra obtener descripciones lógicas e interpretables, lo que es bueno desde una perspectiva de gestión. Ejemplo de esto es que en general los clientes de segmentos más propensos a pasar desde Formato 1 a Formato 2, tienen más edad, una proporción mayor de hombres, menor proporción de solteros y mayor cantidad de hijos. Por otro lado, los clientes más propensos a pasar desde Formato 2 a Formato 1, son en general más jóvenes que el promedio de clientes de Formato 2, con menos hijos y en proporciones de sexo más cercanas.

Otro resultado interesante con respecto a las segmentaciones es que en general a medida que se avanza en los segmentos, también se avanza en el gasto anterior (excepto en Apertura de Tarjeta) pero no necesariamente se logra efectos causales de mayor magnitud. Ejemplo de esto es el caso del cruce de Formato 2 a Formato 1 y la apertura de tarjeta, donde si bien los segmentos obtienen efectos causales distintos, estos no siguen una tendencia clara. Por otro lado, en los segmentos de cruce desde Formato 1 a Formato 2, sí presenta una tendencia al alza a medida que se pertenece a un segmento de mayor propensión. Caso contrario es el que se produce para la primera compra web, donde a medida que se pertenece a un segmento de mayor propensión, el incremental en gasto va disminuyendo, al igual que a la contribución que se reduce prácticamente a la mitad entre el segmento 1 y el segmento 5.

Con respecto a la robustez de los resultados, al variar el número de no activados por cada activado, se obtienen resultados que quedan dentro de los respectivos intervalos

de confianza. Esto habla bien a cerca de la robustez de las estimaciones, sobre todo a cerca de qué tanto podrían variar los resultados si se usan nuevas muestras de clientes. Por otro lado, donde se producen mayores diferencias en los resultados es cuando se cambian los modelos de propensión, en el caso de la primera compra web el resultado escapa levemente de los intervalos de confianza del incremental en gasto y en transacciones, pero en el resto de los hitos se mantiene dentro de los márgenes.

Por el lado del potencial de aplicabilidad de la memoria, se observa que en los hitos evaluados existe un número importante de clientes que podrían ser gestionados. Dado que, solo los clientes que no han comprado en el canal físico, pero no en el digital superan los 4 millones.

Finalmente, la gestión mediante hitos utilizando Propensity Score Matching es útil para la gestión de los clientes con una perspectiva de valor corporativo. Al tener resultados en efectos causales promedios por cada segmento y por cada distinto hito, es posible priorizar clientes a los cuales gestionar y decidir cuales hitos realizar primero, en base a su incremental esperado. También con esto se puede decidir distintos niveles de inversión a la hora de gestionar cada grupo, aplicando una gestión más agresiva para aquellos segmentos e hitos que impliquen un mayor efecto causal.

9.2 Sobre la metodología

Los hitos estudiados en la memoria son en general de carácter poco frecuente: la primer compra web la realiza un 1% de los potenciales clientes en un mes, la apertura de tarjeta un 0.3%, el cruce de Formato 1 a Formato 2 un 3.4% y el cruce de Formato 2 a Formato 1 un 3.3%. Por lo que para la estimación de los puntajes de propensión se realiza un submuestreo de la clase mayoritaria: aquella que no se activa en los hitos en a la ventana observada. Calibrar los modelos así, da como resultado que el mejor balance se logra con una proporción 1:4 entre activados y no activados, dando como resultado que, al aumentar la proporción de no activados, o sea, acercándola a la proporción real se obtiene diferencias más grandes entre las medias de los grupos. Esto resulta interesante al ser contraintuitivo, pues al hacer un submuestreo de las clases, lo que se está obteniendo son probabilidades más “infladas” que las reales.

Por otro lado, en la memoria se obtiene que el modelo que mejor balance es generado por regresión logística o Red Neuronal, quedando Random Forest como el peor modelo para esta labor. Esto a pesar de que Random Forest en prácticamente todos los casos es la mejor herramienta para separar las muestras en puntajes de propensión, tanto en la curva ROC-AUC como la CAP. Una de las características que comparten tanto la Regresión Logística como la Red Neuronal, es que las predicciones se realizan con combinaciones lineales de las variables observables de los clientes que en su última fase pasan por una función sigmoide, a diferencia del Random Forest que las predicciones son el promedio simple de un número determinado de árboles de decisión.

Con respecto a la selección de método de segmentación, las diferencias de balance de variables entre grupos es la menor cuando se utiliza el método de cortes aleatorios. No obstante, el método de segmentación por quintiles de clientes logra un buen balance también, por lo que, por criterios de interpretabilidad y tiempo de ejecución, podría ser una mejor alternativa utilizar el método de quintiles.

Así es como, la metodología se prueba útil para la estimación de efectos causales, pero es necesario definir de una forma gestionable para los negocios las condiciones con las cuales un cliente se considera activado. Por ejemplo, variaciones interesantes podrían ser ver los hitos de primera compra web separado por negocio y ver el efecto en cada uno.

10. TRABAJOS FUTUROS

Existen varias posibles aplicaciones directas de la metodología aplicada en esta memoria. Lo primero es extender el esquema de gestión basado en valor corporativo con un set de hitos posibles más grandes y con variaciones. Por ejemplo, se podría medir por separado el efecto de la primera compra web en cada negocio por separado, como también empezar a evaluar hitos de distinto tipo como hitos de recompra de productos o activaciones en distintas categorías de productos dentro de un mismo negocio.

Otro punto interesante podría ser la evaluación de los hitos en distintas ventanas de tiempo, con la finalidad de ver si es que el efecto causal se produce en ventanas más cortas de tiempo. Se puede realizar un análisis donde un mismo set de clientes matcheado se realiza la evaluación causal en ventanas de semanas hasta llegar a un 1 año, para encontrar el punto en el tiempo en donde el efecto causal se mantiene constante pues ahí será donde la diferencia de los grupos dejó de ser influenciada por el hecho de que el cliente haya realizado el hito.

Como la evaluación de los efectos causales se hace en el pasado, con clientes que ya realizaron el hito, entonces uno de los supuestos más fuertes es que en el futuro los clientes que realicen los hitos también tendrán el mismo efecto causal que aquellos que lo hicieron en el pasado. Otro supuesto fuerte es que el efecto causal es reproducible en clientes gestionados, dado que en el pasado los clientes mayoritariamente fueron autoseleccionados para realizarlos. Todos estos puntos pueden ser probados mediante experimentaciones, por esto resulta importante estimar los efectos causales en ventanas más cortas, porque medir los resultados de una campaña después de 1 año podría resultar arriesgado y poco rentable para el negocio, resulta más viable hacerlo en ventanas de 1 o 3 meses.

Por otro lado, una de las aplicaciones más interesantes para un holding que agrupa varios negocios del retail, como el caso de esta memoria, es la evaluación de efectos de canibalización en el cruce de negocio o canal. Como se mencionó en la justificación de la memoria, algunos de los negocios del holding comparten categorías de productos, por lo que pasar un cliente de un negocio a otro podría producir cierto nivel de canibalización de compras. Esto puede ser probado fácilmente aplicando la metodología con la que se evalúan los efectos causales, incluso tomando el ejercicio ya hecho y cambiar la variable objetivo que será evaluada. Por lo tanto, cada hito puede ser evaluado en las variables que el holding encuentre interesantes, para la canibalización por cruce de negocio se puede usar el gasto antes y después en el negocio de origen y para la canibalización por canal se puede usar el gasto antes y después en el canal de origen.

11. BIBLIOGRAFÍA

1) "CRISP-DM 1.0 Step-by-step data mining guide", SPSS 2000

2) The central role of Propensity Score Matching in Observational Studies for Causal Effects, Paul R. Rosenbaum, Donald B. Rubin, 1983

3) "Propensity Score-Matching methods for non-experimental causal studies", Rajeev H. Dehejia and Sadek W ahba.

4) "Lecture #11: Classification & Logistic: Regression", Weiwei Pan, Pavlos Protopapas, Kevin Rader, Fall 2016.

5) Retail chileno crecería 2,4% anual a 2021, uno de los menores avances en la region" Economia y negocios EMOL http://www.economiaynegocios.cl/noticias/noticias.asp?id=348293

6) "Chile: 5 tendencias que marcarán al retail y consumo masivo chileno" America Retail. https://www.america-retail.com/chile/chile-5-tendencias-que-marcaran-al-retail-y-consumo-masivo-chileno/

7) "Experimental design and Analysis" Chapter 9, Howard Saltnam 2018.

8) "Why Propensity Score Should Not Be Used For Matching" Gary King and Richard Nielsen, 2015

9) Snedecor, George W. and Cochran, William G. (1989), *Statistical Methods*, Eighth Edition, Iowa State University Press

10) Levene, Howard (1960). "Robust tests for equality of variances"

11) "The unequal variance t-test is an underused alternative to Student's t-test and the Mann Whitney U test", Graeme D. Ruxton, 2006

12) "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples" Peter C. Austin, 2009

13) "Prognostic score–based balance measures for propensity score methods in comparative effectiveness research", Elizabeth A. Stuart, Brian K. Lee, and Finbarr P. Leacy. 2013

14) Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. 2001

15) K-means, EM, Gaussian Mixture, Graph Theory, Francis Bach Lecture 3, 2014.

16) “Impact of subsampling and tree depth on random forests”, Roxane Duroux, Erwan Scornet

17) “Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics.” Austin P. C. 2010a

12. ANEXOS

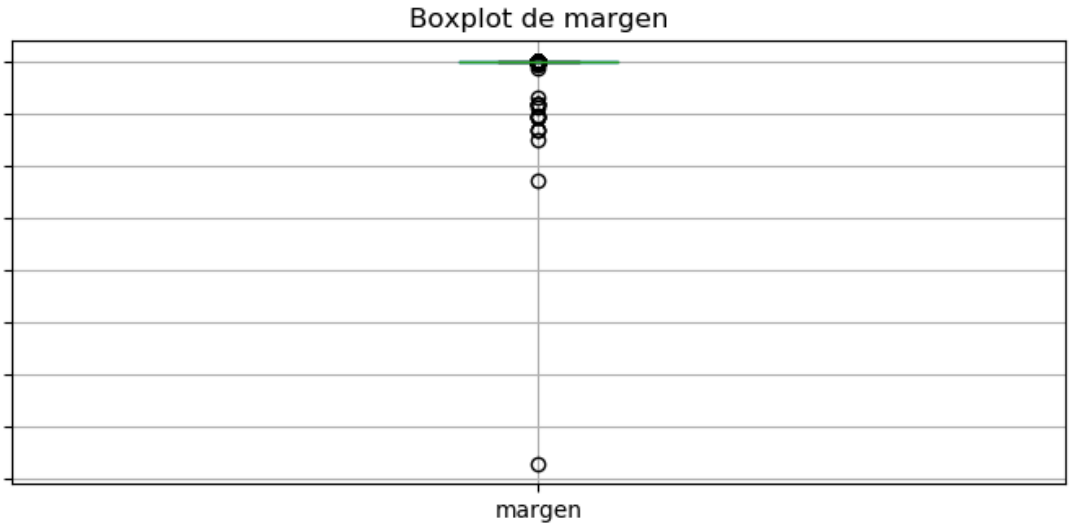
1. Todas las variables

Sociodemográfico	Transaccional
Edad	Gasto en cada negocio
Cantidad hijos	Transacciones en cada negocio
Sexo	Cliente
Estado civil	Recency en cada negocio
Region	Crecimiento 3 meses vs 6 en cada negocio
Tarjeta	Share gasto web en cada negocio
Tenencia tarjeta	Email
Monto cupo	Envios (recibidos)
Tipo tarjeta	Aperturas
	Ratio de apertura

2. Estadísticos caso primera compra web, post eliminación de valores atípicos

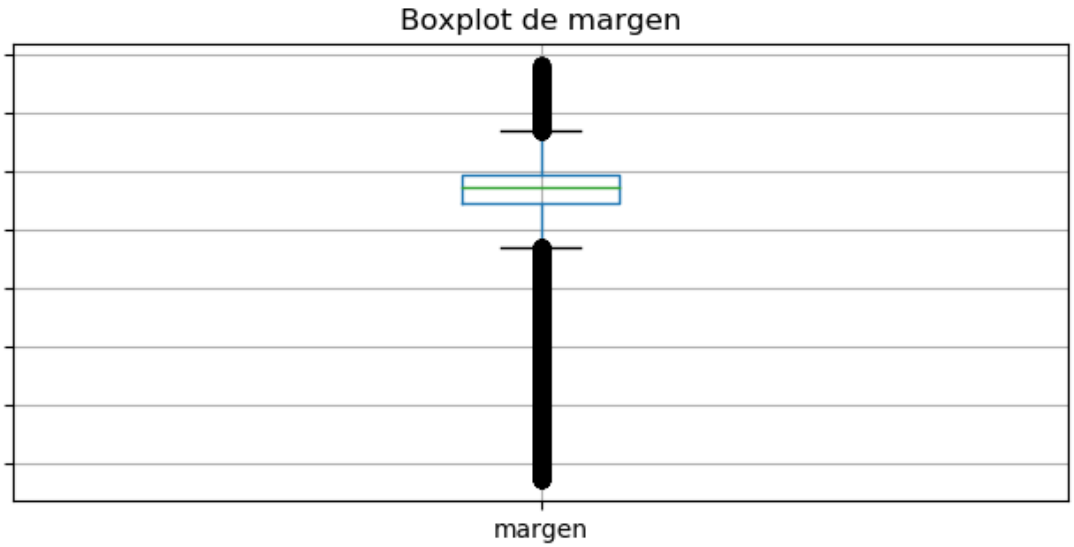
Item	Gasto F1 12 M	Gasto F2 12M	Gasto F3 12M	margen
mean	100.0%	100.0%	100.0%	100.0%
std	141.8%	169.1%	166.9%	0.0%
min	0.7%	0.7%	0.8%	-671.9%
0.01	2.0%	1.3%	1.5%	193.8%
0.1	7.2%	4.7%	6.8%	25.0%
0.25	16.3%	12.1%	16.5%	68.8%
0.5	43.8%	39.6%	42.1%	112.5%
0.75	122.9%	112.8%	108.3%	146.9%
0.9	264.7%	253.0%	241.4%	181.3%
0.99	705.9%	883.2%	907.5%	234.4%
max	1144.4%	1733.6%	1642.9%	440.6%

3. Boxplot de margen sin eliminación de valores atípicos

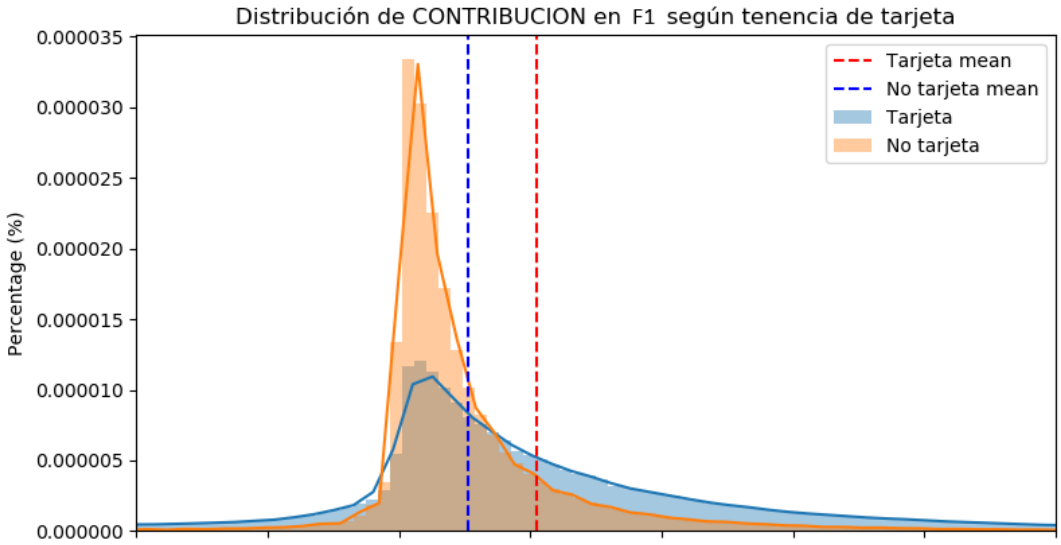


4. Boxplot de contribución con eliminación valores atípicos

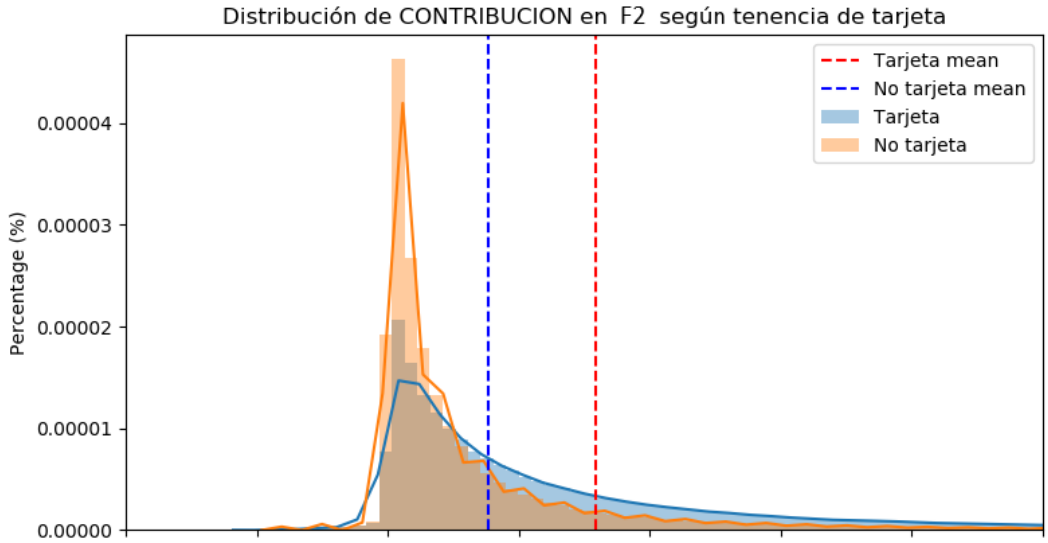
5. Boxplot de margen con eliminación de valores atípicos



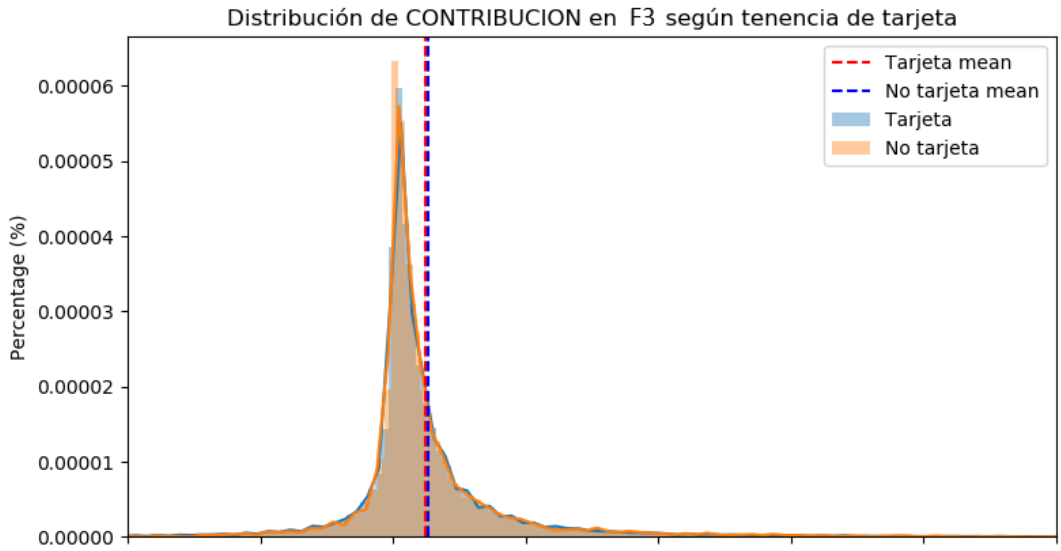
6. Distribución de Contribución en Formato 1, según tenencia de tarjeta



7. Distribución de Contribución en Formato 2 según tenencia de tarjeta

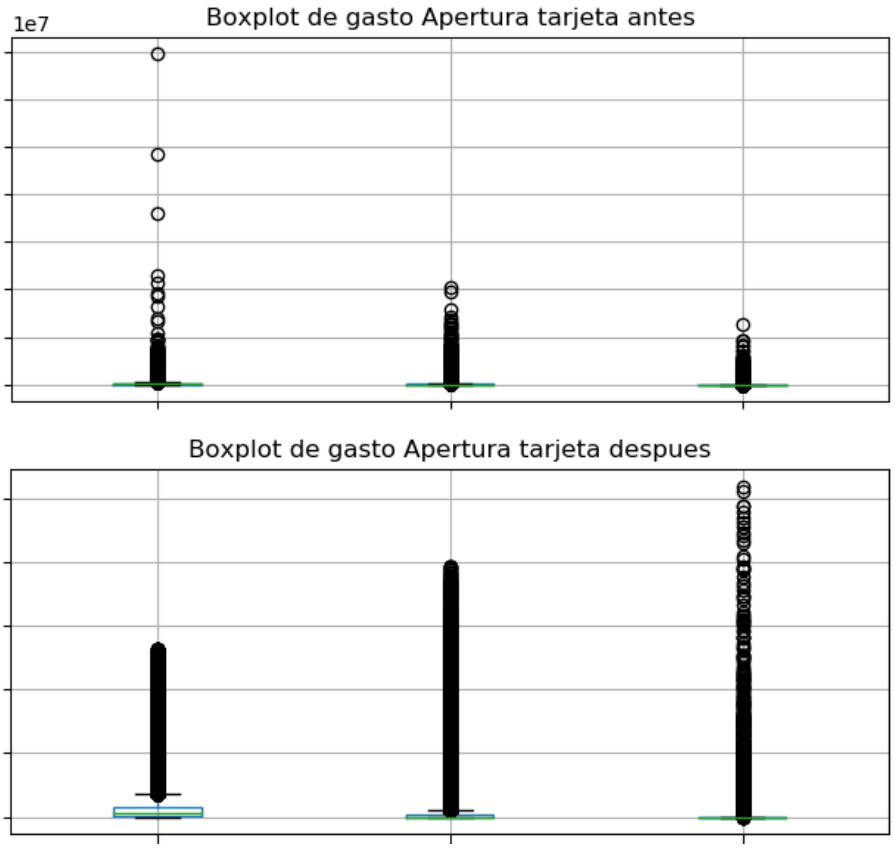


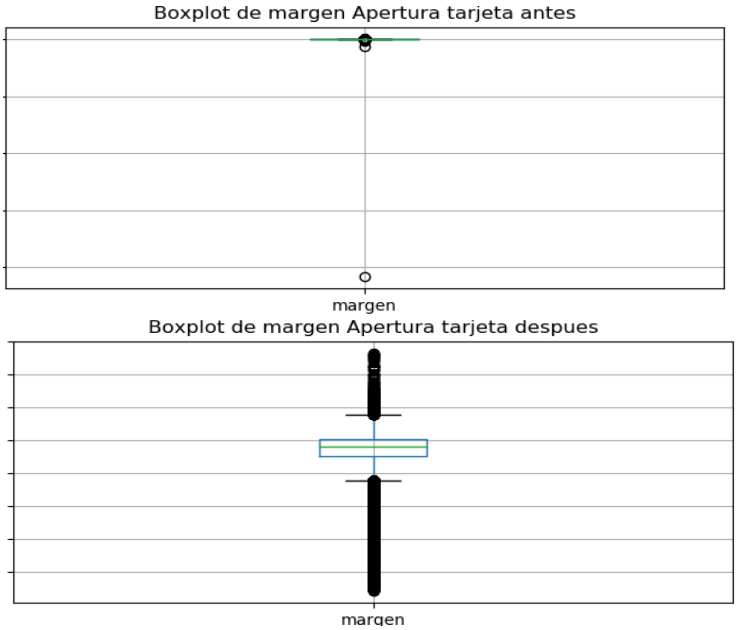
8. Distribución de Contribución en Formato 3 según tenencia de tarjeta



9. Efecto causal en segmentaciones caso primera compra web

10. Boxplot de limpieza de datos apertura de tarjeta

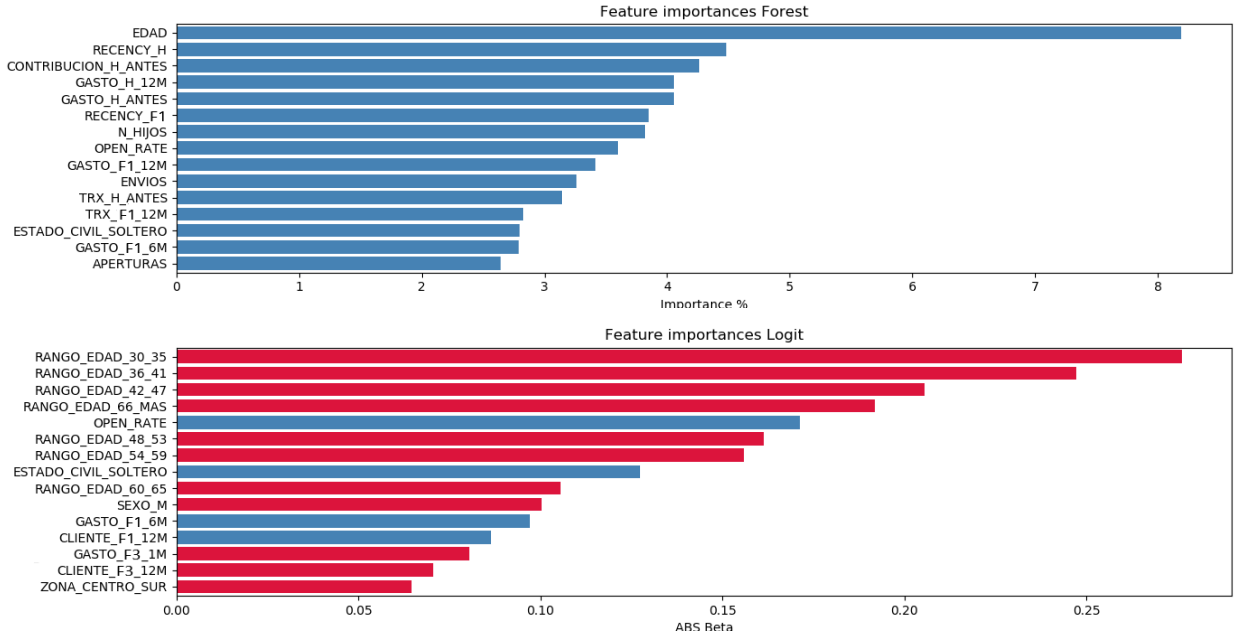




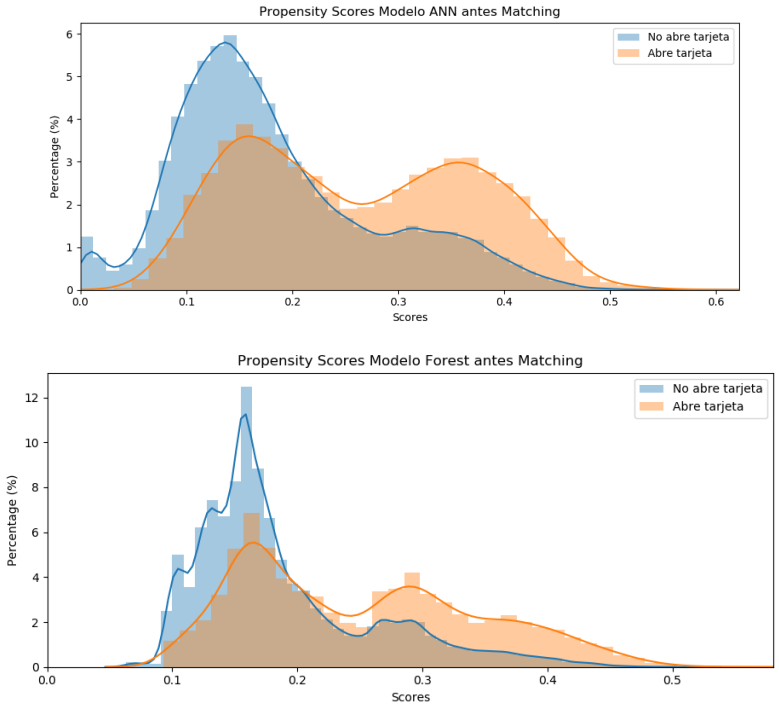
11. Balance de las variables según método de regresión logística en apertura de tarjeta

VARIABLES	SMD antes	SMD después	P Value
GASTO_F1_6M	0.097	0.017	0.185
CONTRIBUCION_H_ANTES	0.007	0.019	0.134
RECENCY_F1	0.210	0.005	0.672
ENVIOS	0.007	0.017	0.179
CLIENTE_F3_12M	0.040	0.014	0.267
SEXO_M	0.150	0.015	0.248
APERTURAS	0.032	0.019	0.134
ESTADO_CIVIL_SOLTERO	0.337	0.025	0.050
RANGO_EDAD_54_59	0.109	0.008	0.550
CLIENTE_F1_12M	0.200	0.002	0.856
EDAD	0.383	0.035	0.006
RECENCY_H	0.114	0.012	0.362
OPEN_RATE	0.152	0.010	0.443
RANGO_EDAD_48_53	0.095	0.018	0.158
GASTO_F1_12M	0.073	0.022	0.084
GASTO_H_ANTES	0.011	0.027	0.030
GASTO_H_12M	0.011	0.027	0.030
TRX_H_ANTES	0.068	0.014	0.278
TRX_F1_12M	0.144	0.014	0.270
RANGO_EDAD_42_47	0.116	0.036	0.005
RANGO_EDAD_36_41	0.136	0.017	0.174
GASTO_F3_1M	0.000	0.017	0.182
ZONA_CENTRO_SUR	0.057	0.004	0.777
RANGO_EDAD_30_35	0.120	0.037	0.003
RANGO_EDAD_60_65	0.075	0.008	0.507
RANGO_EDAD_66_MAS	0.146	0.019	0.140
N_HIJOS	0.300	0.029	0.023

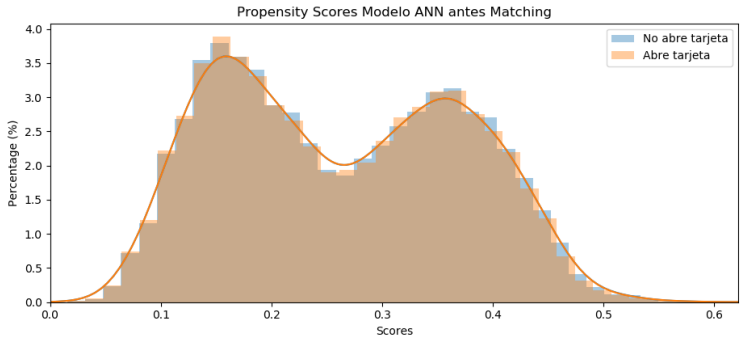
12. Variables más importantes en apertura de tarjeta comercial

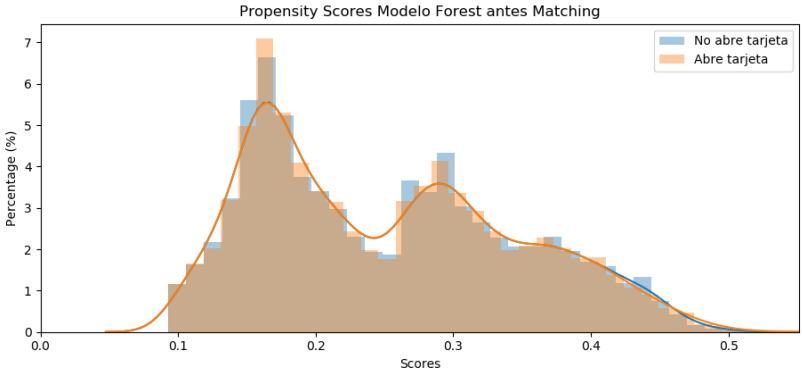


13. Separación de las distribuciones de propensión según modelo, apertura de tarjeta comercial, antes del matching



14. Separación de distribuciones de propensión según modelo, apertura de tarjeta comercial, después del matching



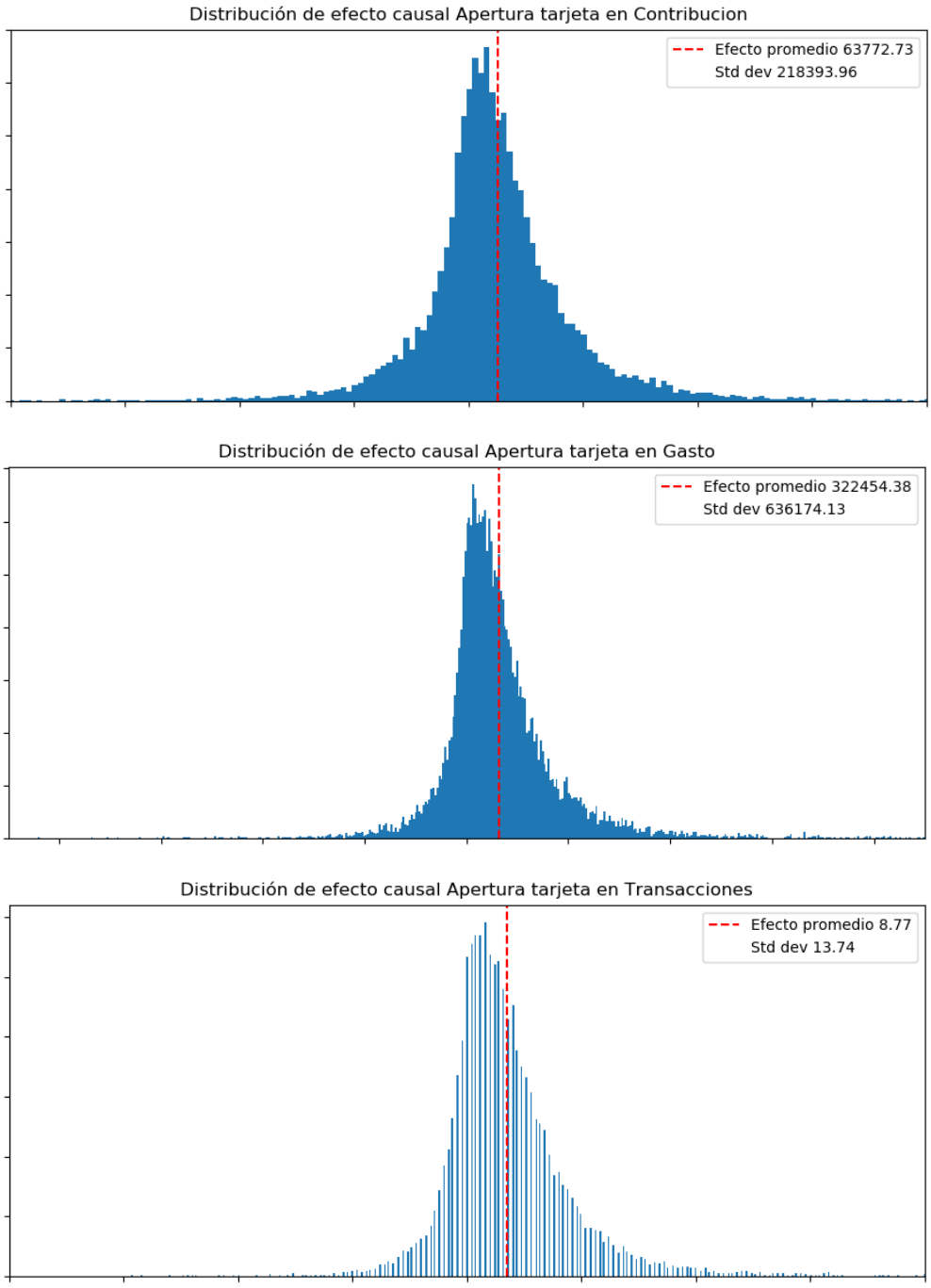


15. Balance en segmentos apertura de tarjeta

Segmento	Kmeans	Quintiles	Random cuts
1	0.022	0.025	0.024
2	0.026	0.026	0.024
3	0.079	0.022	0.018
4	0.020	0.032	0.028
5	0.029	0.028	0.027
Promedio	0.035	0.027	0.024

16. Efecto causal en segmentaciones apertura tarjeta comercial

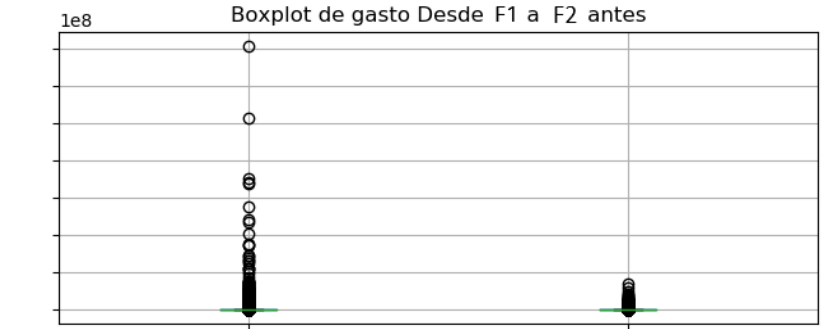
17. Histograma de efectos causales en caso de apertura de tarjeta

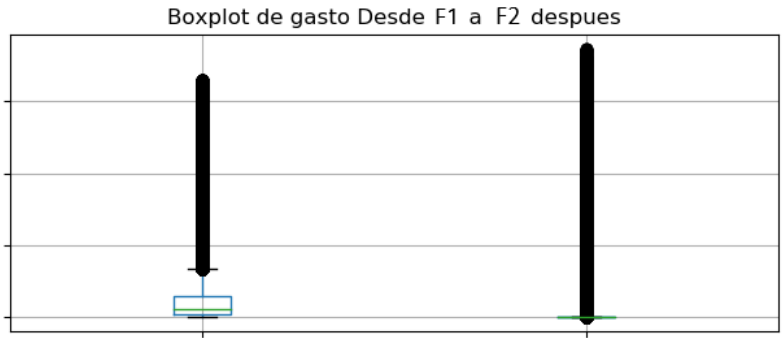


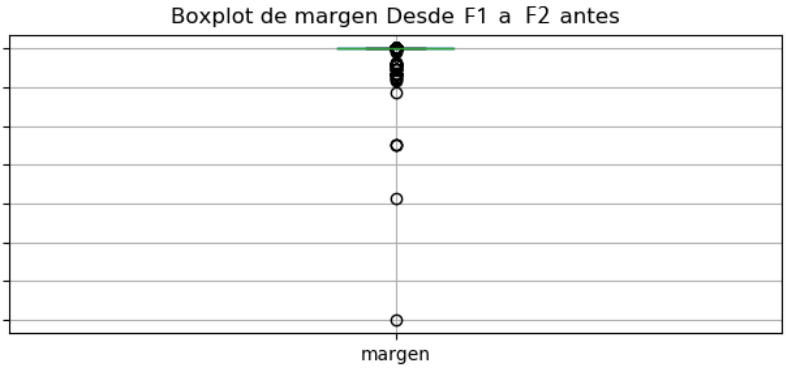
18. Intervalos de confianza apertura de tarjeta en efectos promedio

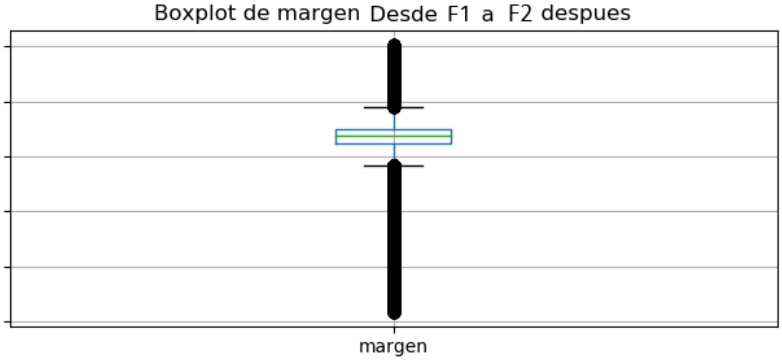
19. Intervalos de confianza apertura de tarjeta en efectos por segmento

20. Boxplots de limpieza de datos, paso de cliente Formato 1 a Formato 2

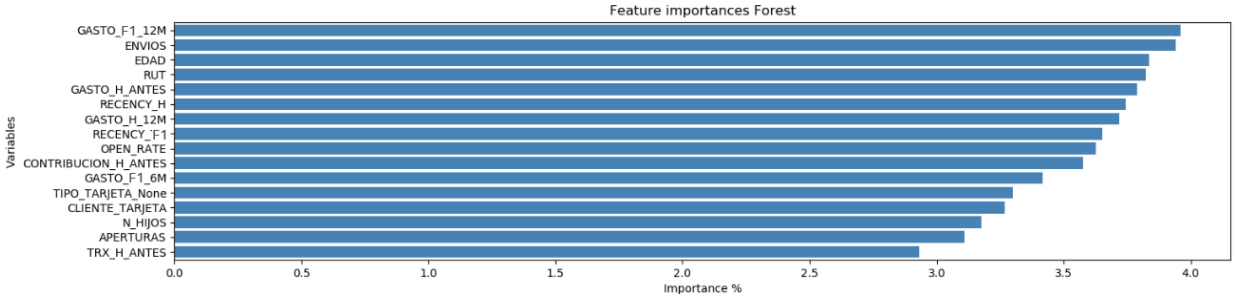


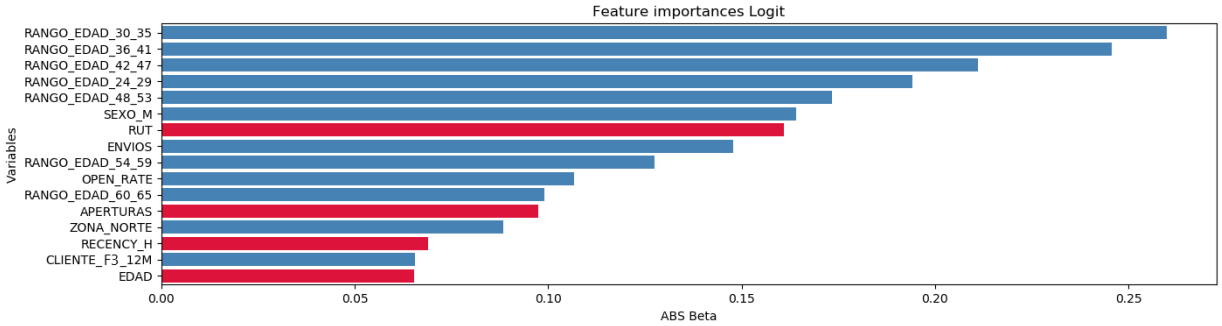






21. Variables más importantes de Formato 1 a Formato 2 y balance entre grupos en cada una de ellas.

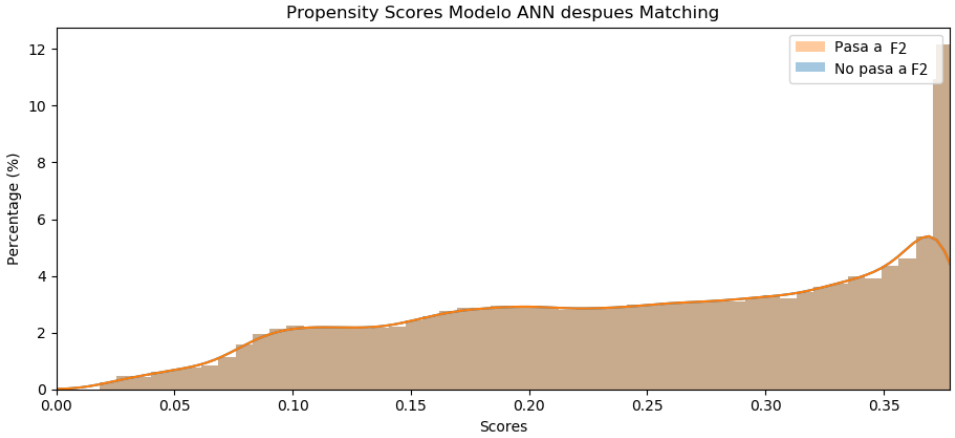


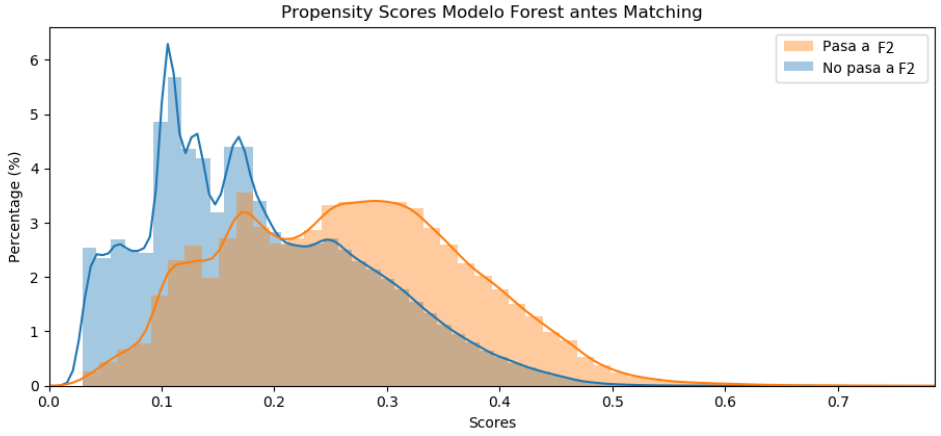


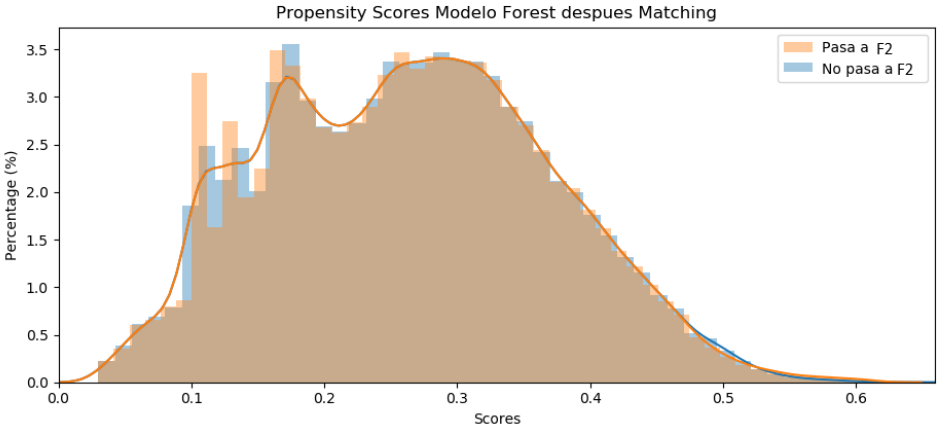
VARIABLES	SMD antes	SMD después	P Value
GASTO_F1_6M	0.328	0.004	0.450
CONTRIBUCION_H_ANTES	0.202	0.010	0.034
RECENCY_F1	0.239	0.002	0.707
ENVIOS	0.333	0.003	0.502
RANGO_EDAD_24_29	0.086	0.001	0.865
CLIENTE_F3_12M	0.249	0.008	0.081
SEXO_M	0.149	0.014	0.004
APERTURAS	0.227	0.007	0.151
RANGO_EDAD_54_59	0.052	0.008	0.097
EDAD	0.181	0.022	0.000
RECENCY_H	0.257	0.002	0.704
OPEN_RATE	0.259	0.002	0.702
RANGO_EDAD_48_53	0.067	0.007	0.139
GASTO_F1_12M	0.357	0.002	0.734
GASTO_H_ANTES	0.369	0.004	0.423
GASTO_H_12M	0.369	0.004	0.423
TRX_H_ANTES	0.284	0.006	0.176
RANGO_EDAD_42_47	0.076	0.018	0.000
RANGO_EDAD_36_41	0.084	0.021	0.000
RUT	0.202	0.018	0.000
RANGO_EDAD_30_35	0.042	0.014	0.004
RANGO_EDAD_60_65	0.045	0.002	0.692
CLIENTE_TARJETA	0.484	0.016	0.001

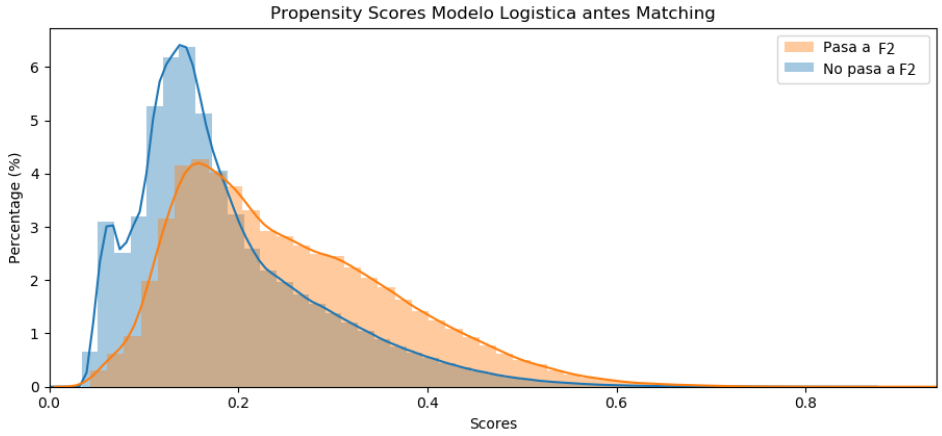
TIPO_TARJETA_None	0.484	0.016	0.001
N_HIJOS	0.131	0.005	0.316
ZONA_NORTE	0.050	0.005	0.271

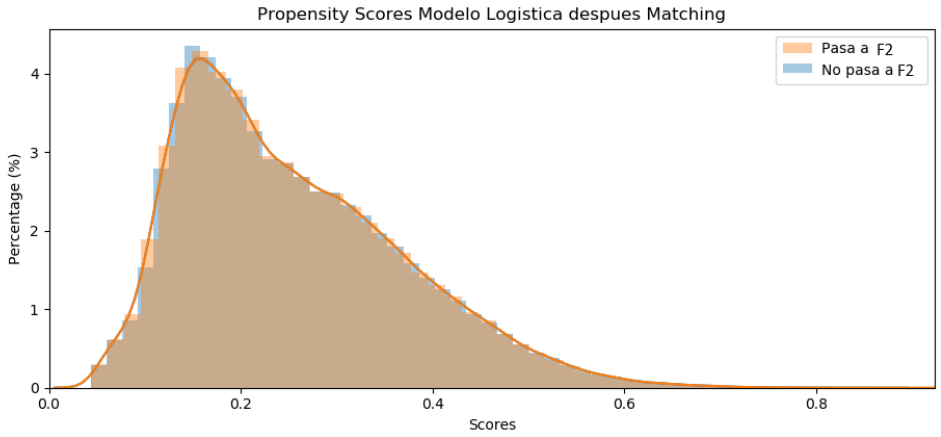
22. Separación de las distribuciones de probabilidad antes y después del match en distintos modelos, caso Formato 1 a Formato 2











23. Balance en segmentos según método de segmentación, cruce de negocio de Formato 1 a Formato 2

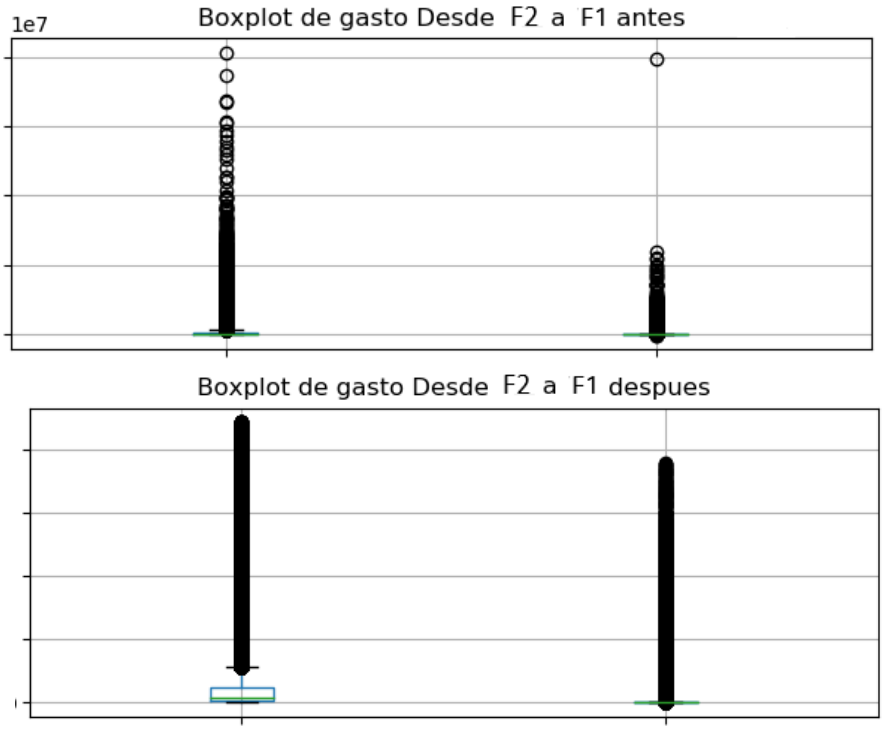
Segmento	Kmeans	Quintiles	Random cuts
1	0.016	0.014	0.012
2	0.014	0.014	0.011
3	0.013	0.011	0.011
4	0.014	0.013	0.011
5	0.011	0.018	0.017
Promedio	0.014	0.014	0.012

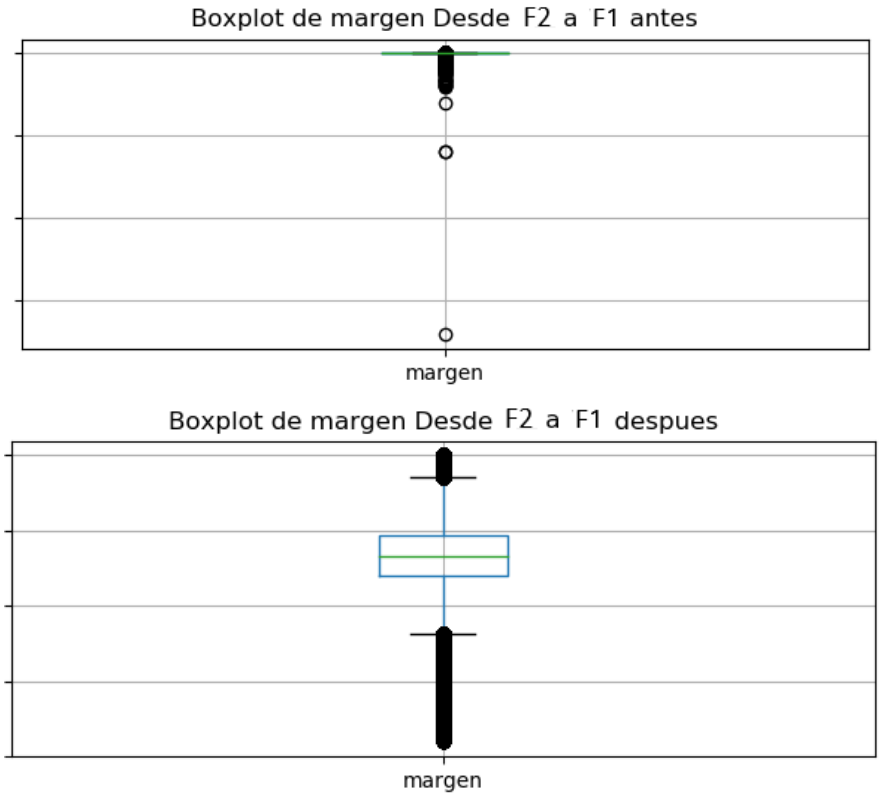
24. Resultado de los efectos causales por segmento, cruce de negocio de Formato 1 a Formato 2

25. Histogramas de efectos causales en los clientes, hito cruce de negocio Formato 1 a Formato 2

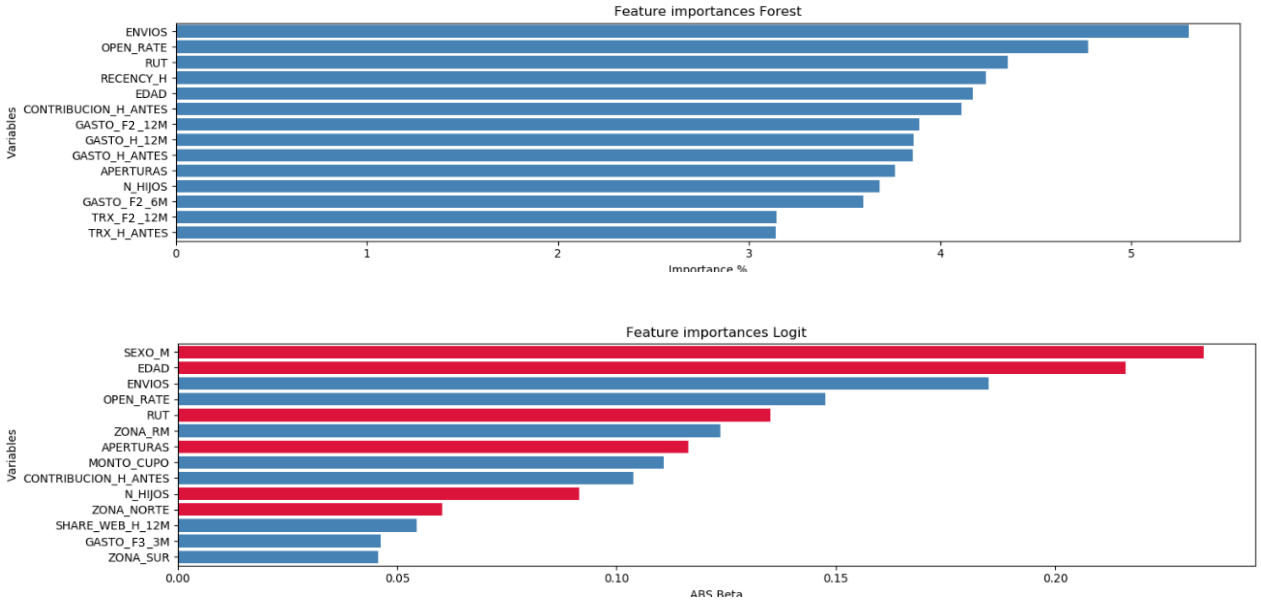
26. Intervalos de confianza y significancia de los resultados de cruce de negocio Formato 1 a Formato 2

27. Boxplots de limpieza de datos, paso de cliente Formato 2 a Formato 1



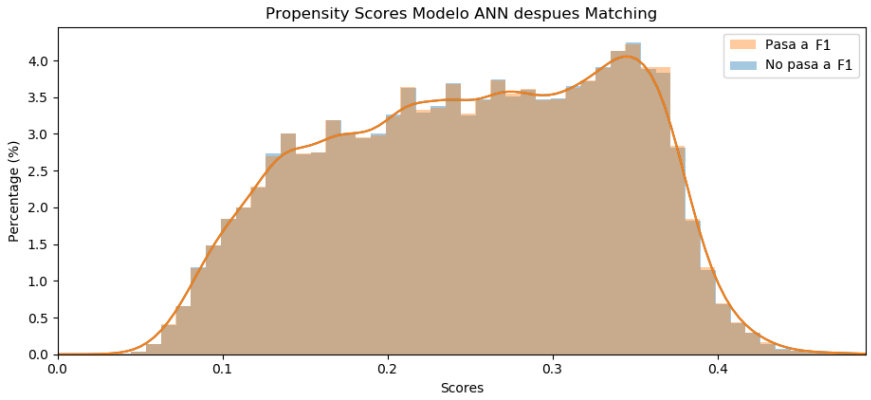


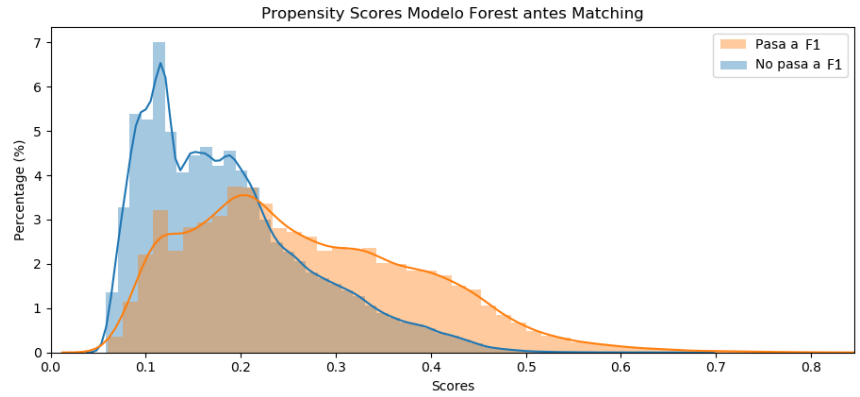
28. Variables más importantes paso de Formato 2 a Formato 1

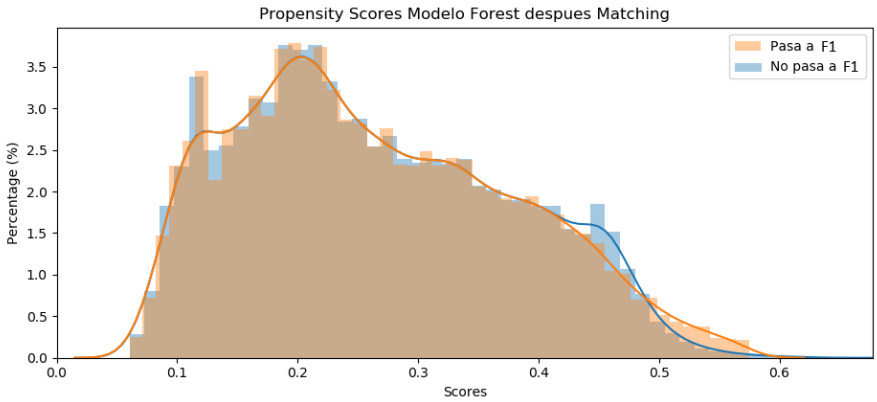


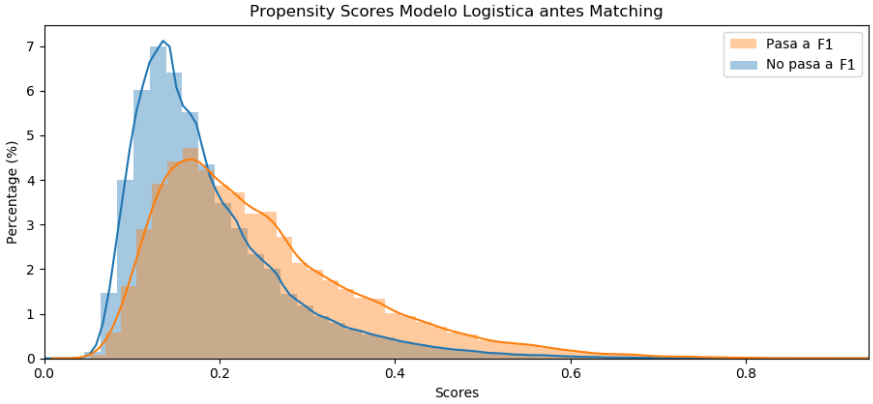
VARIABLES	SMD antes	SMD después	P Value
TRX_F2_12M	0.101	0.004	0.630
GASTO_F3_3M	0.105	0.003	0.716
CONTRIBUCION_H_ANTES	0.154	0.001	0.856
ENVIOS	0.340	0.015	0.050
ZONA_SUR	0.007	0.000	0.964
SEXO_M	0.234	0.008	0.304
APERTURAS	0.222	0.012	0.118
MONTO_CUPO	0.289	0.006	0.467
EDAD	0.121	0.001	0.932
RECENCY_H	0.093	0.006	0.457
OPEN_RATE	0.273	0.021	0.008
ZONA_RM	0.160	0.002	0.764
GASTO_H_12M	0.151	0.001	0.893
GASTO_H_ANTES	0.151	0.001	0.893
TRX_H_ANTES	0.132	0.004	0.636
SHARE_WEB_H_12M	0.149	0.006	0.437
RUT	0.103	0.003	0.711
GASTO_F2_6M	0.100	0.003	0.741
N_HIJOS	0.147	0.012	0.128
ZONA_NORTE	0.107	0.006	0.483
GASTO_F2_12M	0.127	0.002	0.767

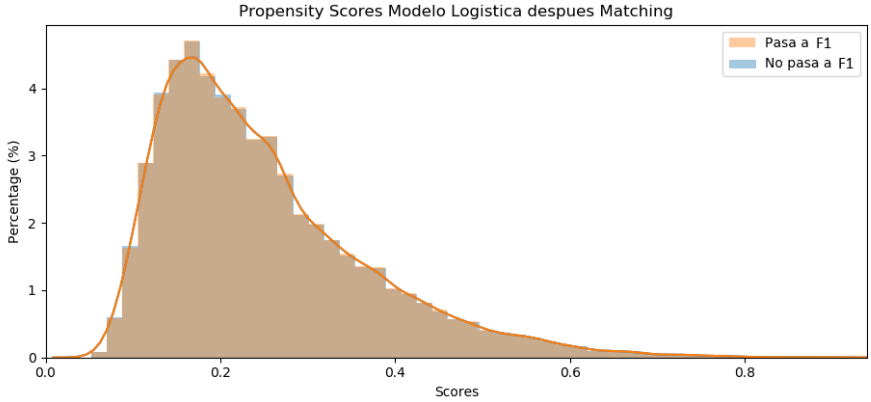
29. Separación de las distribuciones de probabilidad antes y después del match en distintos modelos, caso Formato 2 a Formato 1




Propensity Scores Modelo Forest antes Matching
Pasa a F1
No pasa a F1
Percentage (%)
Scores


Propensity Scores Modelo Forest despues Matching
Pasa a F1
No pasa a F1
Percentage (%)
Scores


Propensity Scores Modelo Logistica antes Matching
Pasa a F1
No pasa a F1
Percentage (%)
Scores

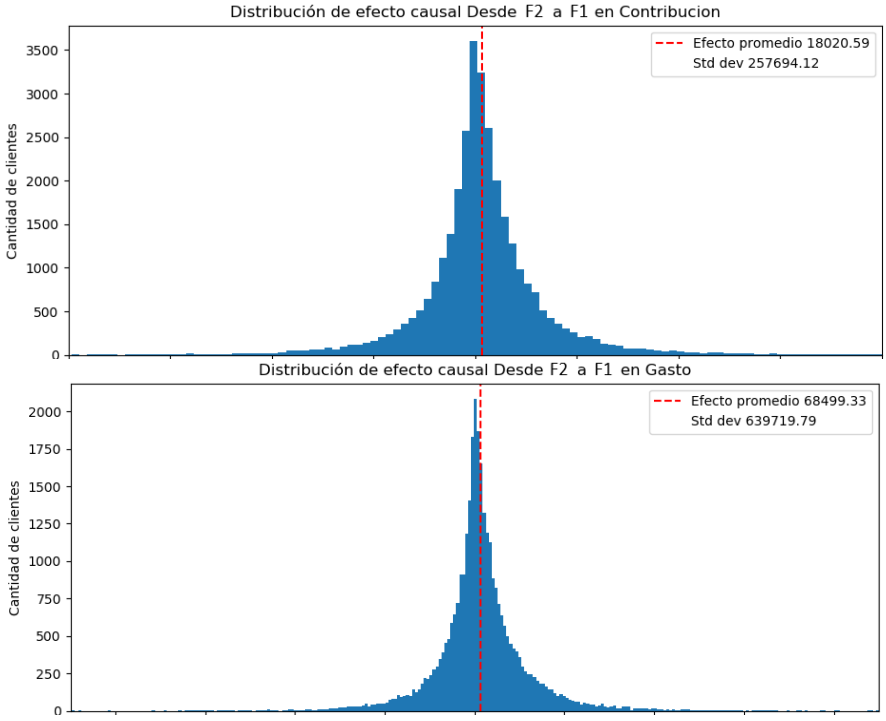

Propensity Scores Modelo Logistica despues Matching
Pasa a F1
No pasa a F1
Percentage (%)
Scores

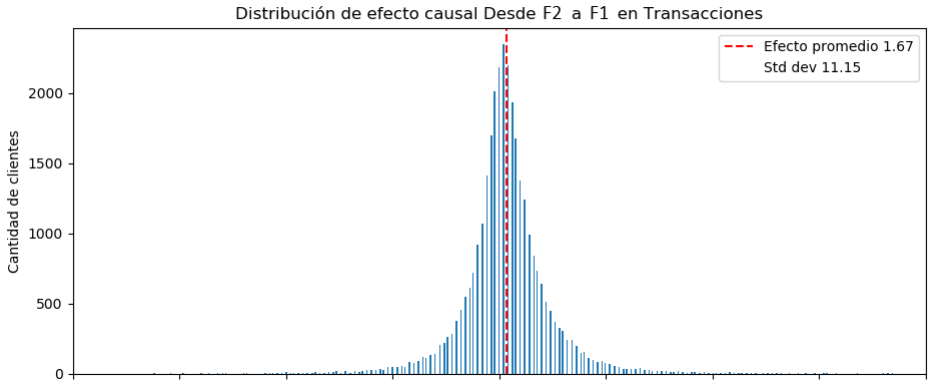
30. Balance en segmentos según método de segmentación, cruce de negocio de Formato 2 a Formato 1

Segmento	Kmeans	Quintiles	Random cuts
1	0.022	0.000	0.015
2	0.016	0.018	0.018
3	0.000	0.020	0.014
4	0.016	0.017	0.015
5	0.016	0.014	0.013
Promedio	0.014	0.014	0.015

31. Resultado de los efectos causales por segmento, cruce de negocio de Formato 2 a Formato 1

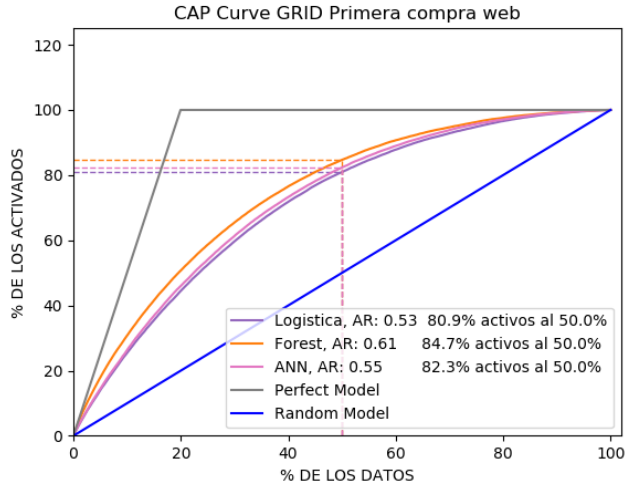
32. Histogramas de efectos causales en los clientes, hito cruce de negocio Formato 2 a Formato 1

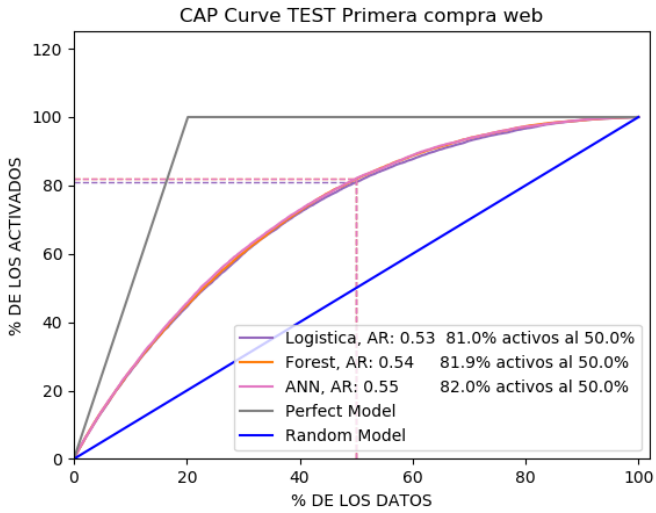




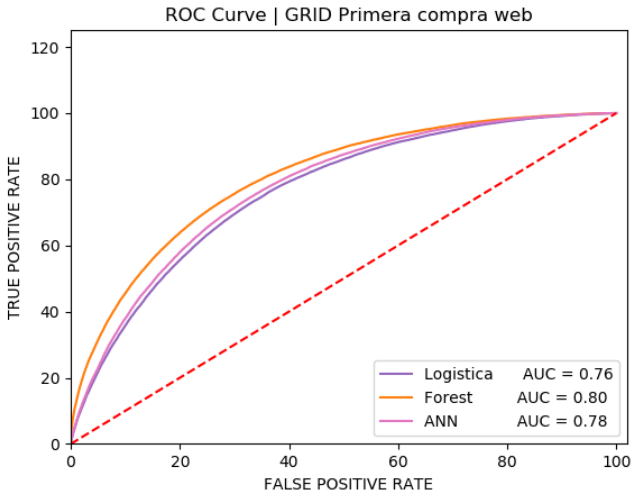
33. Intervalos de confianza y significancia de los resultados de cruce de negocio Formato 2 a Formato 1

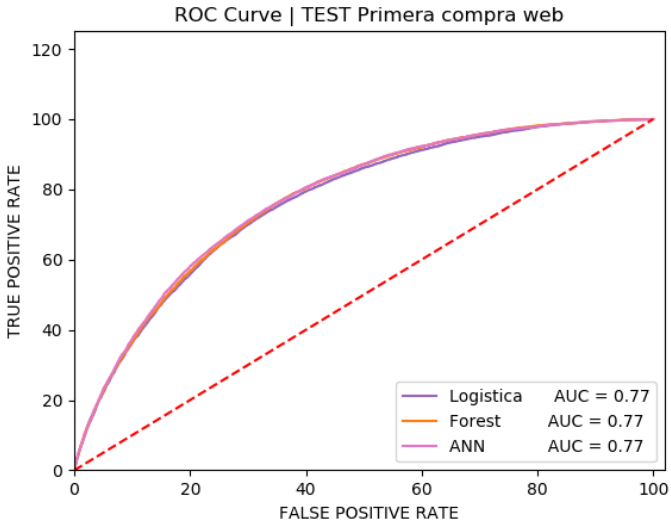
34. Curva cap primera compra web de modelo calibrando en enero 2019.



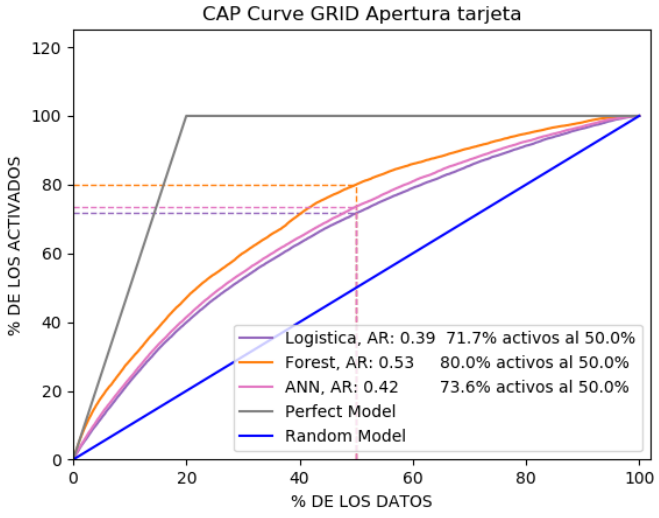


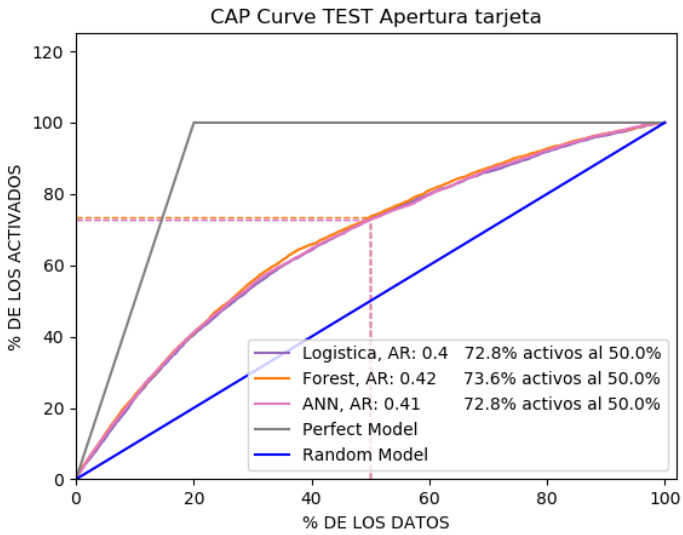
35. ROC-AUC cap primera compra web de modelo calibrando en enero 2019.



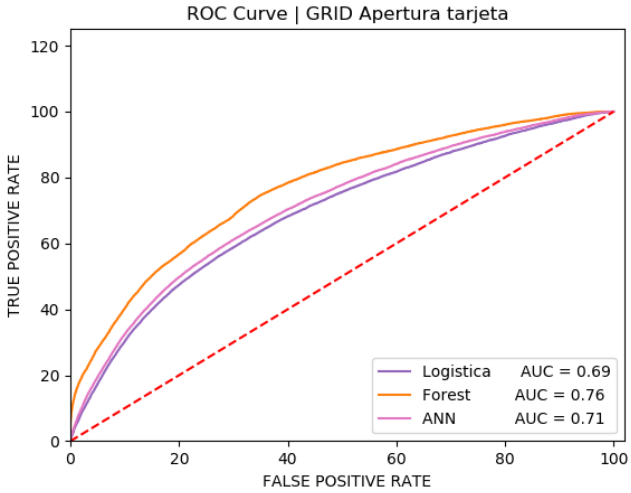


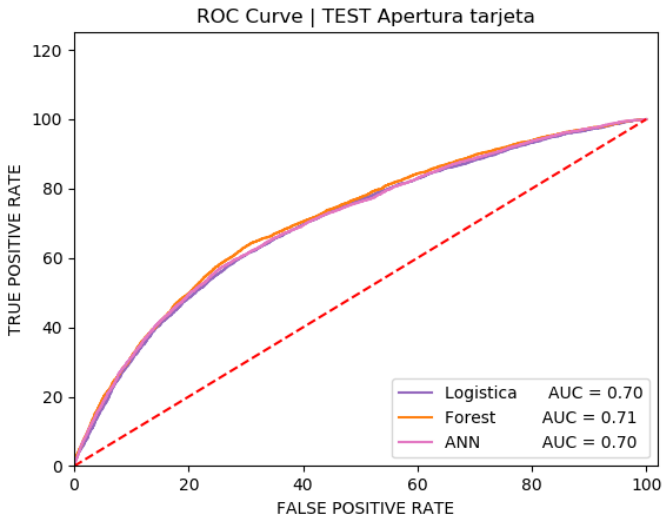
36. Curva CAP apertura de tarjeta modelo calibrado a enero 2019.



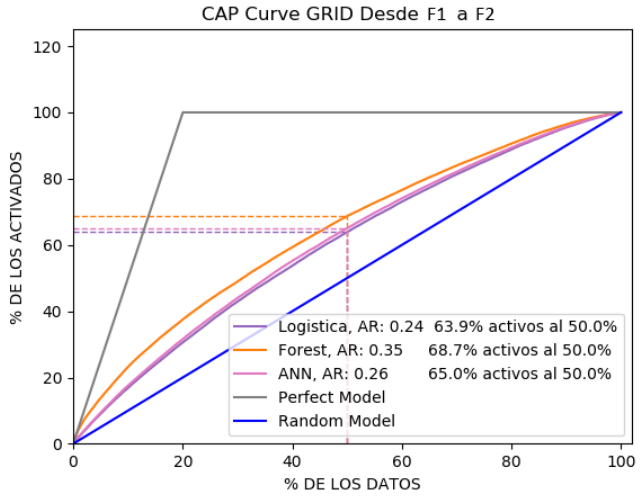


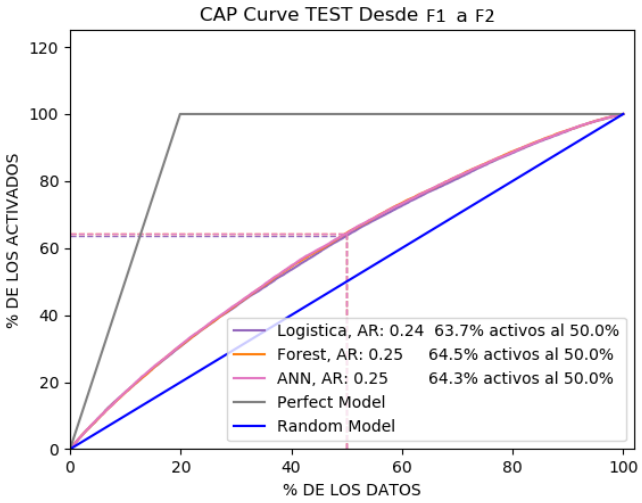
37. Curva ROC-AUC Apertura de tarjeta modelo calibrado a enero 2019



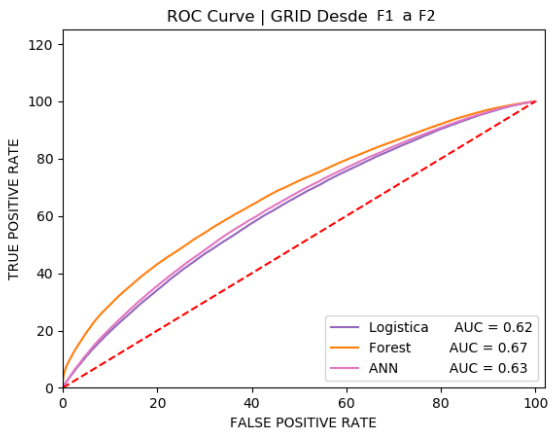


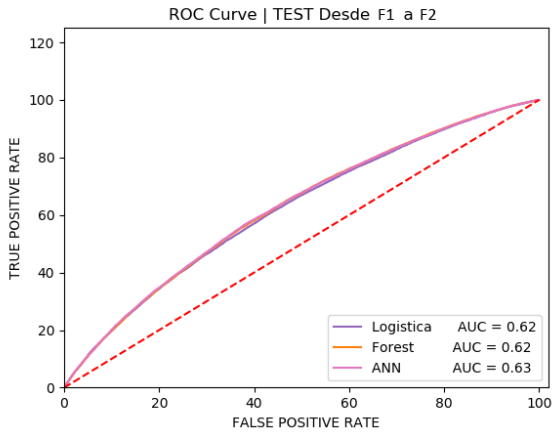
38. Curva CAP cruce de negocio de Formato 1 a Formato 2 calibrado a enero 2019



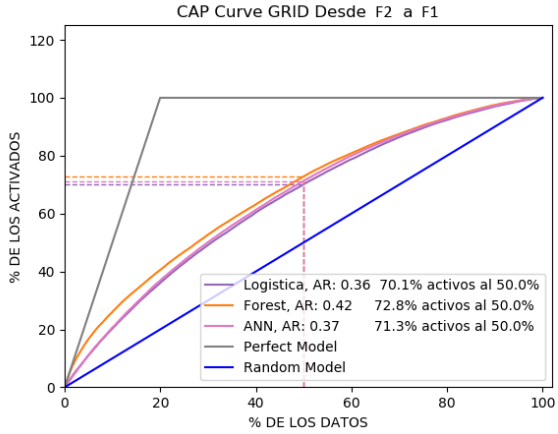


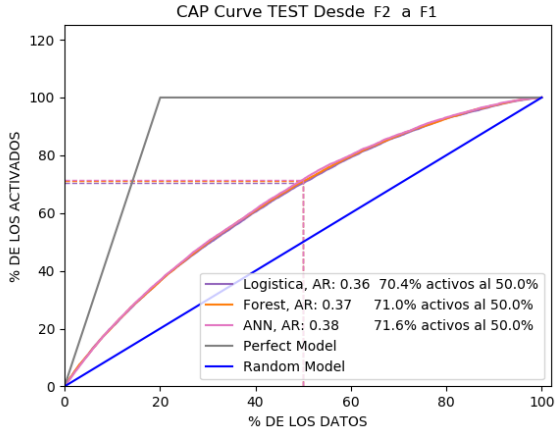
39. ROC-AUC curve cruce de negocio de Formato 1 a Formato 2, calibrado a enero 2019



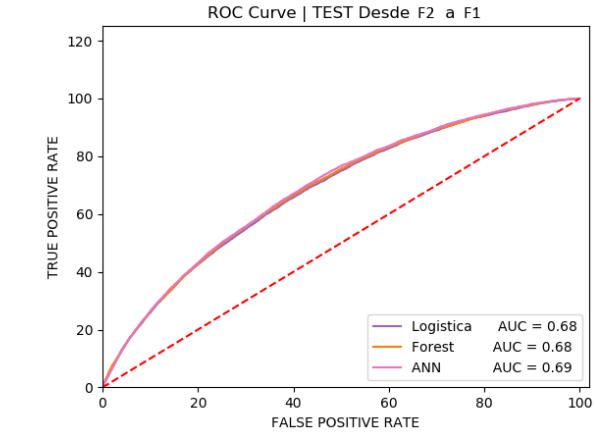


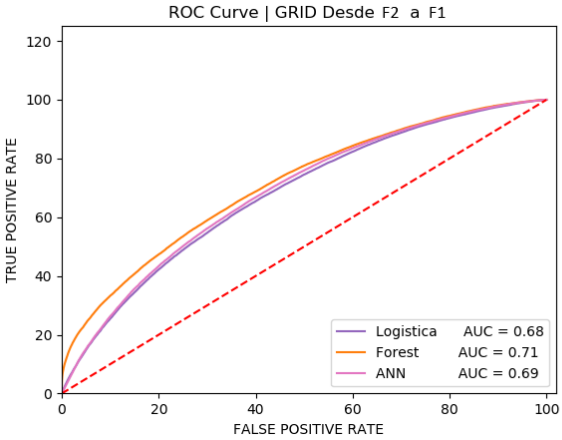
40. Curva CAP cruce de negocio de Formato 2 a Formato 1, calibrado a enero 2019



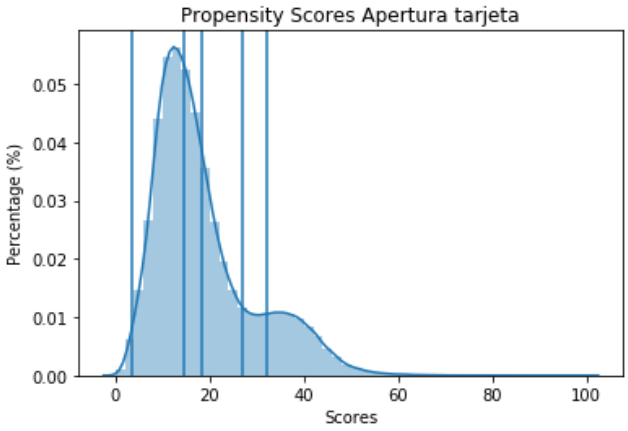


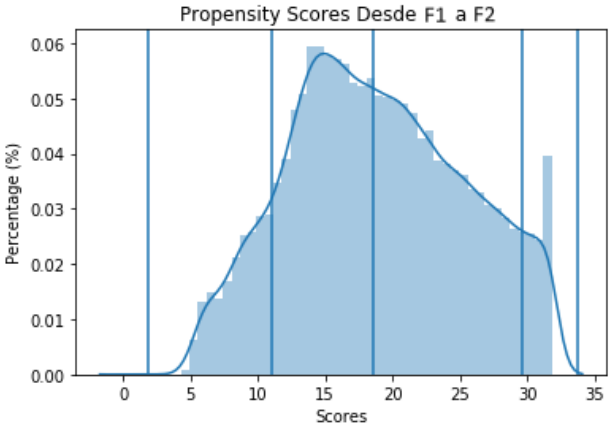
41. Curva ROC cruce de negocio de Formato 2 a Formato 1

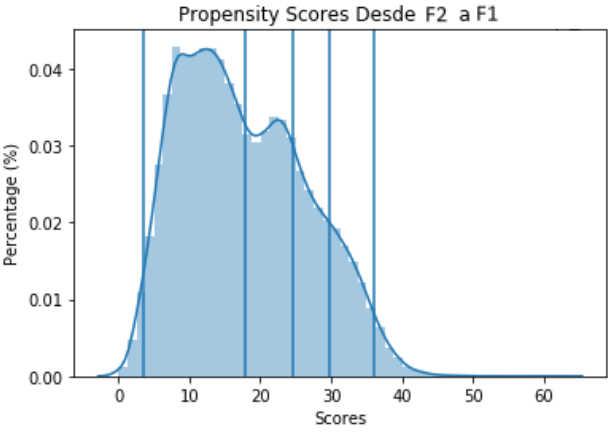




42. Distribución probabilidades de clientes potenciales







43. Análisis potencial de incremento en contribución de gestión de hitos