UNIVERSIDAD TORCUATO DI TELLA

TESIS DE MAESTRÍA

Maestría en Econometría

"Herramientas de Machine Learning Aplicadas al Cálculo de Efectos de Tratamiento en Campañas de Marketing"

Alumna: María Paula Albónico

Legajo: 17M1208

Profesor: Gabriel Martos Venturini

Mayo 2021

Tabla de contenido

ABSTRACT

Al querer estudiar o estimar el efecto causal de una política en cierta variable de interés, el ideal sería comparar el mismo individuo con y sin tratamiento, lo cual, en la práctica resulta, en general, imposible. Una alternativa es llevar a cabo un experimento aleatorio: elegir una muestra a la cual aplicarle el tratamiento y un grupo de control, para luego analizar cómo impactó la política en cuestión a distintos individuos. Este estudio no sólo sirve para evaluar la efectividad del tratamiento o política realizada, sino también para identificar a qué población conviene dirigir una futura política similar, de manera de aumentar su efectividad.

En esta tesis se pretende consolidar bibliografía y teoría sobre efecto de tratamientos, a la vez de aportar un análisis completo al estudio en datos reales relativos a campañas de marketing. Se trabajó con un conjunto de datos publicado por Kevin Hillstrom comúnmente utilizado para probar nuevas metodologías de ML a la predicción de efectos causales para datos reales. Algunos ejemplos son los trabajos de Devriendt et at. [7] y Berrevoets et al. [4][1].

1 Introducción

Estimar los llamados "Efectos de Tratamiento" es de gran utilidad a la hora de la toma de decisiones o de implementación de ciertas intervenciones o tratamientos. Cada tratamiento de algún tipo afecta a los sujetos de dicho tratamiento de diferentes posibles maneras. Por ejemplo, en el campo de la medicina, conocer los efectos de ciertos medicamentos permite sugerir tratamientos médicos "personalizados" según las características y necesidades de cada paciente. Estimar estos efectos personalizados también es importante al lanzar una campaña, por ejemplo política o de marketing. Para que una campaña sea más efectiva, es necesario caracterizar o identificar qué individuos serán más receptivos a cierto mensaje o publicidad. Por otro lado, estimar efectos causales también sirve para evaluar la efectividad de ciertas políticas implementadas por un gobierno.

En cualquiera de los casos anteriores, el estudio del efecto de una (o varias) intervenciones apunta a responder preguntas como si el tratamiento fue efectivo o no, o en el caso de que haya habido más de una variante, cuál fue mejor. ¿Se puede modificar de alguna manera el mensaje o la población a la cual dirigirlo para aumentar la efectividad? Las conclusiones de dicho análisis permiten mejorar los resultados de la intervención en cuestión.

La estimación de efectos de tratamiento tiene un gran obstáculo: se tiene información faltante. Puntualmente, para medir cómo influye cierto tratamiento en una variable de interés sería necesario conocer el impacto tanto en la población tratada como en la de control. Sin embargo, cada individuo sólo pertenece a uno de los grupos y se desconoce cómo hubiera reaccionado si pertenecía al grupo complementario.

El estudio de inferencia contra factual en datos reales (a diferencia de datos sintéticamente generados) ha sido un tópico de interés en economía, estadística, salud, medicamentos, colocación de publicidad, entre otras áreas de estudio. Los datasets reales más comúnmente usados en la bibliografía como marco de comparación para modelos de Machine Learning son el IHDP (*Infant Health and Development Program*) del trabajo de Gross [9], *Jobs* (donde se estima el efecto de capacitación en el estado de empleo de los individuos) de LaLonde [12], *Twins* (basado en el nacimiento de mellizos en Estados Unidos entre 1989 y 1991) del trabajo de Louzos [14], y un dataset publicado por Kevin Hillstrom[2]. En esta tesis se trabajó con este último conjunto de

[1] Dado que en general en estos trabajos se estudió esencialmente el uplift para 'visit' en la campaña femenina, sólo se comparará en ese caso usando la métrica numérica en común, el valor del Qini.
[2] El dataset, publicado por la consultora de marketing MineThatData, se encuentra disponible en el siguiente link: https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

datos, explorado en detalle en la sección 3, que contiene información sobre dos campañas de marketing realizadas por un negocio de indumentaria. Concretamente, la población se divide en tres grupos disjuntos: el grupo de control y los grupos tratados con una u otra campaña. Claramente, el dueño de la empresa quiere la solución a la siguiente pregunta: ¿cómo se puede aumentar las ventas y las ganancias? El objetivo del análisis es, entonces, interpretar a partir de los datos si las campañas fueron efectivas o no, si alguna fue mejor que la otra, y si, por ejemplo, se puede identificar en qué individuos tuvieron más impacto para dirigir futuras campañas sólo a aquellos individuos que el análisis identificó como más propensos a reaccionar positivamente. Para esto, se estimó el efecto causal con distintas metodologías, comparando algunas tradicionales como regresión lineal con otras técnicas más actuales de Machine Learning como KNN (por sus siglas en inglés "K Nearest Kneighbours") [10] y Random Forests para efectos de tratamiento, como Honest Causal Random Forests [19]. Las distintas metodologías consideradas, junto con los resultados obtenidos y su comparación, se desarrolla en la sección 4.

A la hora de comparar distintas metodologías de estimación de efectos causales aparece un nuevo obstáculo, o en realidad es el mismo de siempre: no se cuenta con un "verdadero" efecto de tratamiento para comparar con la predicción. Existen varias métricas sin embargo, que sirven para evaluar la performance de un modelo de efectos causales, las cuales se explican en la sección 2.1.

2 Generalidades sobre Efectos de Tratamiento

En esta sección se pretende sintetizar el marco teórico que se usó a lo largo de la tesis sobre efectos de tratamiento, así como también introducir las definiciones y notaciones necesarias.

Se va a suponer que para cada individuo o unidad $i = 1, \dots n$ se tiene la siguiente información:

- $X_i \in \mathbb{R}^d$ (vector de atributos, observables)
- $T_i \in \{0; 1\}$ (indicador de tratamiento)
- Y_i (variable de interés, "*outcome*")

Se trabajó en el marco teórico de *potential outcomes* propuesto por Rubin (1974) [18], también llamado RBC (por sus siglas en inglés "*Rubin Causal Model*"). Se asume que, previo al tratamiento, cada unidad tiene "por naturaleza" dos posibles valores de Y asignados: $\big(Y_i(0), Y_i(1)\big) \in \mathbb{R}^2$. Sin embargo, para cada individuo sólo se observa uno de estos valores (resultado factual), dependiendo el tratamiento, y no se conoce cuál sería el valor con el tratamiento contario (resultado contra factual). De ahí el nombre de "resultados potenciales". Concretamente, una vez aplicado el tratamiento, sólo se conoce

$$Y_i^{obs} \;=\; T_i \;\; Y_i(1) \;+\; (1 - T_i)\, Y_i(0) \;=\; Y_i(T_i) \qquad (1)$$

En general, si se nota por $\mathcal{D}_i \;=\; (Y_i(0), Y_i(1), X_i)$ a las características del individuo brindadas por la naturaleza, se asume que $(\mathcal{D}_i)_{i=1}^n$ son independientes e idénticamente distribuidas (iid).

Interesa entonces cuantificar el efecto de cierta política o tratamiento: $Y(1) - Y(0)$. Este impacto o incremento, es el denominado *uplift*. Por ejemplo, algunos parámetros de interés pueden ser:

- ITE (*Individual Treatment Effect*): $ITE_i \;=\; Y_i(1) - Y_i(0)$
- CATE (*Conditional Average Treatment Effect*): $CATE(x) \;=\; \mathbb{E}[Y(1) \;-\; Y(0) \mid X = x]$
- ATE (*Average Treatment Effect*): $ATE \;=\; \mathbb{E}[Y(1) \;-\; Y(0)]$

Claramente, el ITE es el parámetro ideal que se desearía conocer: el efecto para cada individuo. Por otro lado, el CATE es un poco más general, mirando conjuntamente individuos con las mismas características. Todavía más general aún, el ATE resume la información de los parámetros anteriores. Notar que ATE se puede obtener a partir de CATE:

$$ATE \;=\; \mathbb{E}[CATE(X) \mid X]$$

Notar además que en el caso de que la variable de interés sea dicotómica, la esperanza de cada *potential outcome* se reemplaza (ya que es equivalente) por la probabilidad.

En cualquier caso, al tratar de estimar las métricas definidas anteriormente, aparece lo que Holland (1986) [11] describe como "el problema fundamental de la inferencia causal": la imposibilidad de conocer ambos *potential outcomes*. Para superar este obstáculo, los métodos de estimación generalmente consideran los siguientes supuestos [18]:

- **Supuesto 1:** SUTVA, por sus siglas en inglés, "*Stable Unit Treatment Value Assumption*", asume que los resultados potenciales de cada unidad no varían en función del tratamiento asignado a otras unidades, y que en definitiva la probabilidad de que el individuo i pertenezca al grupo de tratamiento ($T_i = 1$) sólo depende de $\mathcal{D}_i$. SUTVA implica, por ejemplo, que dos individuos con los mismos atributos tienen la misma probabilidad de ser tratados.

- **Supuesto 2**: Positividad (también llamado *Overlap*). Las probabilidades de tratamiento son no triviales. Matemáticamente: $0 < \mathbb{P}(T = t \mid X) < 1$. Es decir, para cada valor de $X = x$ hay observaciones tanto en el grupo de control como en el de tratamiento. Este supuesto es importante ya que si, por ejemplo, todos los individuos con cierto conjunto de atributos no hubieran recibido el tratamiento no habría manera de conocer o inferir que hubiera pasado si hubieran sido tratados.
- **Supuesto 3**: Consistencia. El resultado potencial de tratamiento $T = t$ equivale al valor observado si el verdadero tratamiento fue $T = t$. Es decir, se cumple la ecuación (1) ($Y_i^{obs} = Y(T_i)$).

Para poder estimar el efecto de tratamiento a partir de los datos observables, es necesario otro supuesto introducido por Rosenbaum y Rubin (1983) [17]:

- **Supuesto 4:** *Unconfoundedness* (Independencia Condicional) Para cada X, los valores potenciales son independientes de la asignación del tratamiento: $T \perp (Y(0), Y(1)) \mid X$. Heurísticamente, este supuesto garantiza que dado ciertos atributos X, las diferencias entre el grupo de tratamiento y el de control son completamente aleatorias. Aunque es difícil probar que se satisface este supuesto, se va a cumplir en experimentos donde la asignación del tratamiento es aleatoria, ya sea completamente o condicional a ciertos atributos.

Ante la imposibilidad de evaluar el efecto comparando el mismo individuo con y sin tratamiento, estos supuestos permiten estimar el CATE como la diferencia entre el valor esperado de Y de dos grupos similares, habiendo asignado el tratamiento a sólo uno de ellos. Formalmente:

Teorema 1: Dados los supuestos anteriores, se tiene que

$$CATE(x) = \mu_1(x) - \mu_0(x) \quad \forall\, x$$

Siendo $\mu_t(x) = \mathbb{E}[Y^{obs} \mid X = x,\ T = t]$.

La demostración es muy sencilla, y se muestra a continuación a modo ilustrativo, ya que se logra ver la necesidad de los supuestos.

Demostración:

Por definición, se tiene que $CATE(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$. Como se está condicionando a X, por el supuesto 4 (*Unconfoundedness)*, se puede también condicionar por el tratamiento manteniendo la igualdad. Razonando de esa manera para cada término por separado ($t = 0$ y $t = 1$) se tiene:

$$\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y(t) \mid X = x,\ T = t] = \mathbb{E}[Y^{obs} \mid X = x,\ T = t] = \mu_t(x)$$

Además, notar que por el supuesto de *Overlapping* (supuesto 2), ambas esperanzas son calculables. □

La intuición detrás de este teorema es, esencialmente, que por ejemplo, $\mathbb{E}[Y(1) \mid X = x,\ T = 0]$ al ser desconocido se puede aproximar por $\mathbb{E}[Y(1) \mid X = x,\ T = 1]$. Más aún, no es necesaria toda la información de X para que valga esto. Para formalizar esta idea, es necesario introducir una definición.

Definición 1: Definimos el *propensity score* como la probabilidad de que un individuo pertenezca al grupo de tratamiento, dadas sus características x.

$$e(x) \coloneqq \mathbb{P}(T = 1 \mid X = x) = \mathbb{E}(T \mid X = x) \qquad (2)$$

Esta función resume la información de X que es necesaria para la independencia de $(Y(0), Y(1))$ con T. Formalmente:

Teorema 2 (Rosenbaum y Rubin (1983) [17]): Bajo el supuesto de *Unconfoundedness,* se tiene que

$$T \perp (Y(0), Y(1)) \mid e(X)$$

Mirando el ATE, se puede determinar la efectividad de una política a grandes rasgos. Pero para poder analizar el tratamiento más en profundidad y tomar decisiones a partir de los resultados obtenidos es necesario mirar el efecto individual o, en su defecto, el CATE.

En la sección 4 se presentan algunas técnicas en la estimación de efectos de tratamiento, algunas más tradicionales, y otras más nuevas y superadoras. Antes de eso, en el capítulo siguiente, se presenta el problema a resolver y se introducen los datos de estudio.

2.1. Métricas de comparación de modelos

En los modelos tradicionales, las medidas de performance dependen principalmente en la comparación del valor predicho vs el valor real de interés, para cada individuo. Sin embargo, en los modelos de efectos de tratamiento con datos reales, esto es imposible debido al "problema fundamental de la inferencia causal" descripto por Holland (1986) [11], desarrollado anteriormente en esta sección: no se conoce el valor real del efecto de tratamiento.

Existen varías métricas para evaluar la performance de modelos de inferencia causal, dependiendo también del tipo de datos. Por ejemplo, en su paper Wager y Athey (2017) [19] usan el MSE ("*Mean Squared Error*") como métrica para comparar modelos, dado que los autores trabajaron con simulaciones. En estos casos donde se cuenta con los valores conocidos del efecto de tratamiento y/o de ambos *potential outcomes*, generados con alguna distribución elegida, se dispone de otras métricas como ε_{PEHE} ("*Precision in Estimation of Heterogeneous Effect*"), curvas Qini y AUUC. Estas últimas dos métricas también sirven en el caso de datos reales y se explican a continuación, entre otras.

2.1.1 MSE para Efectos de tratamiento:

Se refiere al error cuadrático medio calculado a partir del "Transformed Outcome Loss" (TOL)

A continuación se desarrolla en detalle esta métrica, ya que se usó a lo largo de este trabajo de dos maneras, o con dos objetivos distintos.

1) Comparación de modelos.
2) Selección de modelos o elección de hiperparámetros.

Para definirla, primero es necesario introducir el concepto de *Transformed Outcome*, introducido por Athey e Imbens (2015) [1]. Tanto para *outcomes* binarios como numéricos, se define Y^* en términos del *outcome* observado y el *propensity score* (2), de la siguiente manera.

$$Y_i^* := \frac{T_i - e(x_i)}{e(x_i)\big(1 - e(x_i)\big)} \cdot Y_i^{obs} \tag{3}$$

Notar que en el caso de un experimento completamente aleatorio, como de hecho se podrá asumir en el caso de los datos utilizados (ver sección 3.2), el *propensity score* es constante y no depende de las características X, simplificando la expresión.

Para entender por qué sirve Y^*, y cómo está relacionada con la comparación de modelos, se presentan las siguientes dos propiedades.

Propiedad 1: Bajo los supuestos de *Unconfoundedness* y *Overlapping*, Y^* es un estimador insesgado del CATE.

Demo:

Reescribiendo la expresión de Y^* y tomando esperanza se tiene:

$$\begin{aligned}
\mathbb{E}[Y^*|X] &= \mathbb{E}\left[T \cdot \frac{Y(1)}{e(X)} - (1-T) \cdot \frac{Y(0)}{1-e(X)} \middle| X\right] \\
&= \mathbb{E}\left[T \cdot \frac{Y(1)}{e(X)} \middle| X\right] - \mathbb{E}\left[(1-T) \cdot \frac{Y(0)}{1-e(X)} \middle| X\right] \\
\overline{\overline{U}} \quad & \mathbb{E}[T|X] \cdot \frac{\mathbb{E}[Y(1)|X]}{e(X)} - (1 - \mathbb{E}[T|X]) \cdot \frac{\mathbb{E}[Y(0)|X]}{1-e(X)} \\
&= e(x) \cdot \frac{\mathbb{E}[Y(1)|X]}{e(X)} - \big(1-e(x)\big) \cdot \frac{\mathbb{E}[Y(0)|X]}{1-e(X)} \\
&= \mathbb{E}[Y(1) - Y(0)|X] = CATE(X)
\end{aligned}$$

El supuesto de *Unconfoundedness* se traduce en que la esperanza se distribuye en el producto de variables independientes. El supuesto de *Overlapping* se asume para que para cada X haya individuos tanto en el grupo de control como de tratamiento y por lo tanto las esperanzas tengan sentido. □

Entonces, por esta propiedad, se puede escribir $Y^* = CATE(X) + \nu$, con $\mathbb{E}[\nu|X] = 0$. Con esta notación, se obtiene la siguiente propiedad adicional.

Propiedad 2: Bajo los supuestos de *Unconfoundedness* y *Overlapping*, se tiene que para un estimador $\widehat{CATE}(X)$ vale que:

$$\mathbb{E}\left[\left(Y_i^* - \widehat{CATE}(x_i)\right)^2\right] = \mathbb{E}\left[\left(CATE(x_i) - \widehat{CATE}(x_i)\right)^2\right] + \mathbb{E}[\nu^2] \qquad (4)$$

Demo:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i^* - \widehat{CATE}(x_i)\right)^2 \mid X = x_i\right] &= \mathbb{E}\left[\left(CATE(x_i) + \nu_i - \widehat{CATE}(x_i)\right)^2 \mid X = x_i\right] \\
&= \mathbb{E}\left[\left(CATE(x_i) - \widehat{CATE}(x_i)\right)^2 + 2 \cdot \left(CATE(x_i) - \widehat{CATE}(x_i)\right) \cdot v_i + v_i^2 \mid X = x_i\right] \\
&= \mathbb{E}\left[\left(CATE(x_i) - \widehat{CATE}(x_i)\right)^2 \mid X = x_i\right] \\
&\quad + 2 \cdot \left(CATE(x_i) - \widehat{CATE}(x_i)\right) \cdot \underbrace{\mathbb{E}[v_i \mid X = x_i]}_{=0} + \mathbb{E}[v_i^2 | X = x_i] \\
&= \mathbb{E}\left[\left(CATE(x_i) - \widehat{CATE}(x_i)\right)^2 \mid X = x_i\right] + \mathbb{E}[v_i^2 | X = x_i]
\end{aligned}$$

Por lo tanto, por la ley de expectativas iteradas:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i^* - \widehat{CATE}(x_i)\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(Y_i^* - \widehat{CATE}(x_i)\right)^2 \mid X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(CATE(X) - \widehat{CATE}(x_i)(x_i)\right)^2 \mid X\right]\right] + \mathbb{E}\left[\mathbb{E}[v_i^2|X]\right] \\
&= \mathbb{E}\left[\left(CATE(X) - \widehat{CATE}(x_i)\right)^2\right] + \mathbb{E}[v_i^2]
\end{aligned}$$

□

La conclusión importante de este resultado es que permite descomponer el error de tal manera que demuestra que minimizar el error cuadrático medio es equivalente a minimizar el "*Transformed Outcome Loss*" (TOL) (término de la izquierda).

En resumen, cuando se comparan dos modelos, ya sea de distintas metodologías (KNN, Random Forests, etc) o dentro de una misma metodología para elegir hiperparámetros (por ejemplo, a la hora de elegir el número K óptimo de vecinos más cercanos), se va a preferir que la siguiente expresión sea chica.

$$TOL = \frac{1}{\#obs}\sum_{i}\left(Y_i^* - \widehat{CATE}(X_i)\right)^2$$

2.1.2 Curvas Qini

Radcliffe [15] desarrolla varias métricas numéricas y visuales, a partir de una generalización de los Gain chart para el caso de modelos de *uplift*, llamada curvas Qini.

Para construir dicho gráfico, la población se ordena según un score (cada uno de los modelos a considerar), ordenando primero los individuos identificados como mejores individuos; es decir, se ordena de mejor a peor. En el eje x se indica la población receptora de la campaña, mientras que en el eje vertical se mide el incremento acumulado en la variable *outcome* para cada segmento de población considerado, según cada valor del eje x. Por ejemplo, en el caso de una variable binaria, para cada segmento se compara la proporción de $y = 1$ en el grupo de los tratados $(\#1\ en\ T\ /\ \#T)$ con la proporción de $y = 1$ en el grupo de control $(\#1\ en\ C\ /\ \#C)$ el correspondiente punto de ordenada es:

$$u\ =\ (\#1\ en\ T)\ -\ \frac{(\#1\ en\ C)\cdot(\#\ T)}{\#\ C}$$

La calidad del modelo estudiado determina la forma de la curva: cuanto más se aleja la curva de la diagonal correspondiente a un tratamiento random, más útil es el modelo. A diferencia de las curvas *Gain*, las curvas Qini no son monotónicas, sino que son cóncavas negativas. Esta forma se debe a la posibilidad de un "efecto negativo" de la campaña que esencialmente se resume en que al dirigir la campaña a menos gente, se puede lograr una mayor proporción de éxitos[3]. Después de todo, en este efecto negativo radica la importancia de seleccionar apropiadamente los destinatarios. Ejemplos de curvas Qini para los modelos estudiados en esta tesis son gráficos en Figura 4-20 y Figura 4-21, que se analizan más adelante, en la sección 4.4. Para más detalle teórico sobre las curvas Qini, referirse a la explicación de Figura 2 en Radcliffe [15] para el caso de *outcome* binario, y Figura 4 para el caso continuo.

Para calcular las curvas Qini en nuestros modelos, se utilizó la función "get_qini" de la librería CausalML de Python [19], la cual sigue la definición utilizada por Radcliffe.

2.1.3 Qini score

Para la siguiente métrica también se utilizó la misma función de Python, consistente con la definición del coeficiente Qini según Radcliffe [15].

El coeficiente Qini es una generalización del coeficiente Gini, para el caso de modelos de *uplift*. En el caso de outcomes binarios, este coeficiente se define como el ratio entre el área debajo de la curva Qini para el modelo y por encima de la diagonal, y el área entre la curva Qini óptima y la diagonal. En el caso de un *outcome* continuo, por la imposibilidad de una correcta definición de la curva óptima, el coeficiente Qini se define como el área entre la curva Qini y la diagonal, dividido por la mitad del cuadrado del total de observaciones. Este valor se interpreta como el incremento de ganancias por cabeza.

El valor teórico de Q pertenece al intervalo [-1,1], y cuanto más cercano a 1, mejor. Pero cabe resaltar que al trabajar con aproximaciones del verdadero *uplift*, es posible encontrarse con valores ligeramente fuera de dicho rango.

[3] La noción de éxito está atada a la variable outcome de interés.

2.1.4 AUUC (*area under the uplift curve*)

De manera similar a las curvas Qini, las curvas *uplift* se definen como la diferencia en la variable *outcome* entre los grupos de control y tratamiento, en función del porcentaje de población seleccionada. Luego, se define el AUUC como el valor del área debajo de dicha curva. Cuanto mayor es esta métrica, mejor el modelo. Esta métrica, además, es preferible cuando se trata de datos con *outcome* desbalanceado [14].

Para calcular el AUUC en los modelos considerados, se utilizó la función "auuc_score" de la librería CausalML de Python [19]. Cabe mencionar que se utilizó la opción de normalizar los valores del eje y, para obtener valores que sean comparables entre distintas muestras, por ejemplo entre muestra de entrenamiento y testeo.

3 Datos y análisis preliminar

Como se comentó anteriormente, se trabajó con un conjunto de datos publicado por la consultora de marketing MineThatData en el año 2008.

Cabe mencionar que los datos de Kevin Hillstrom, dueño de la consultora, es popular a la hora de aplicar y probar nuevas metodologías de ML a la predicción de efectos causales para datos reales. Algunos ejemplos son los trabajos de Devriendt et at. [7] y Berrevoets et al. [4][4].

3.1. La campaña

El dataset contiene 64.000 observaciones correspondientes a clientes de una tienda de indumentaria que hicieron alguna compra en los últimos 12 meses. El estudio de marketing que se llevó a cabo es el siguiente:

- Se envió un correo electrónico relacionado con mercadería masculina a 1/3 de los clientes, elegidos aleatoriamente.
- Se envió un correo electrónico relacionado con mercadería femenina a 1/3 de los clientes, elegidos aleatoriamente.
- Al tercio restante de la población no se le envió ningún correo (grupo de control).

Al cabo de dos semanas posteriores a dicha campaña de mails, se registraron los resultados en términos de visitas al sitio de internet, compras realizadas y monto de dichas compras.

Cabe resaltar que se puede considerar que hay dos opciones en lo que respecta a definir la "campaña" y el "tratamiento":

1) Se puede considerar que hay dos campañas independientes, una de mercadería masculina y otra de mercadería femenina, donde el grupo de control siempre es el mismo. En este caso, mitad de la población es tratada (para cada campaña individualmente) y la otra mitad es el grupo de control. Es decir, en cada campaña se excluyen de los datos a los clientes que recibieron al correo correspondiente a la campaña contraria.

2) Se puede considerar que hay una sola campaña, donde el tratamiento en cuestión es haber recibido (algún) correo electrónico. En este caso, se trabaja con el total de las observaciones, donde 2/3 de la población fue tratada y 1/3 no lo fue.

Claro que si a partir de una campaña de marketing se busca saber si funcionó o no, y cómo se puede mejorar para aumentar su efectividad, será importante preservar la información correspondiente a las dos campañas de correo electrónico, y por eso en esta tesis se optó por la primera opción. Además, de esta manera, el trabajo sirve como análisis dos experimentos distintos, y en última instancia se podrá evaluar si hubo diferencias en las conclusiones en las dos campañas, y a qué se debe.

La tabla a continuación lista todas las variables disponibles, junto con su descripción. Cabe mencionar que se comprobó que ninguna de las variables tiene valores faltantes.

Tabla 3-1: Lista de variables:

Atributos históricos:		
Recency	Numérica	Meses desde la última compra
History	Continua	Monto gastado en el último año ($)
history_segment	Categórica	Categoría de gasto

[4] Dado que en general en estos trabajos se estudió esencialmente el *uplift* para 'visit' en la campaña femenina, sólo se comparará en ese caso usando la métrica numérica en común, el valor del Qini.

Mens	Dummy	Dummy que indica si compró mercadería masculina en último año.
Womens	Dummy	Dummy que indica si compró mercadería femenina en último año.
zip_code	Categórica	Clasificación regional: “Urban”, “Suburban” o “Rural”
Newbie	Dummy	Dummy que indica si se trata de un cliente nuevo en los últimos doce meses.
Channel	Categórica	Describe el tipo de canal usado por el cliente en el último año: “Phone”, “Web” o “Multichannel”.
Campaña:		
Segment	Categórica	Categoría del tratamiento: “Mens E-Mail”, “Womens E-Mail” o “No E-Mail”.
Comportamiento post-campaña (*outcomes* de interés):		
Visit	Dummy	Dummy que indica si el cliente visitó el sitio en las siguientes dos semanas
Conversion	Dummy	Dummy que indica si el cliente efectuó alguna compra en las siguientes dos semanas
Spend	Continua	Monto gastado en las siguientes dos semanas ($)

La tabla de frecuencias a continuación confirma las proporciones de cada tratamiento (aproximadamente 1/3 de la población para cada opción).

Tabla 3-2: Tabla de frecuencias para ‘segment’

	Frecuencia (#)	**Frecuencias relativas (%)**
Womens E-Mail	21.387	33,42%
Mens E-Mail	21.307	33,29%
No E-Mail	21.306	33,29%
Total	64.000	100%

También es necesario corroborar la aleatoriedad del tratamiento. La importancia de esto se verá cuando se analice la validez de los supuestos (ver sección 3.3).

3.2. Independencia del tratamiento

Una manera de evaluar la aleatoriedad del tratamiento es viendo las proporciones de cada tratamiento en submuestras aleatorias. Como se ve en la tabla a continuación, y mirando conjuntamente con las frecuencias en la población total (Tabla 3-2), la proporción de población asignada a cada tratamiento es cercana al 33% y sin predominancia de ningún segmento en particular, lo que sugiere que la asignación del tratamiento fue, efectivamente, aleatoria.

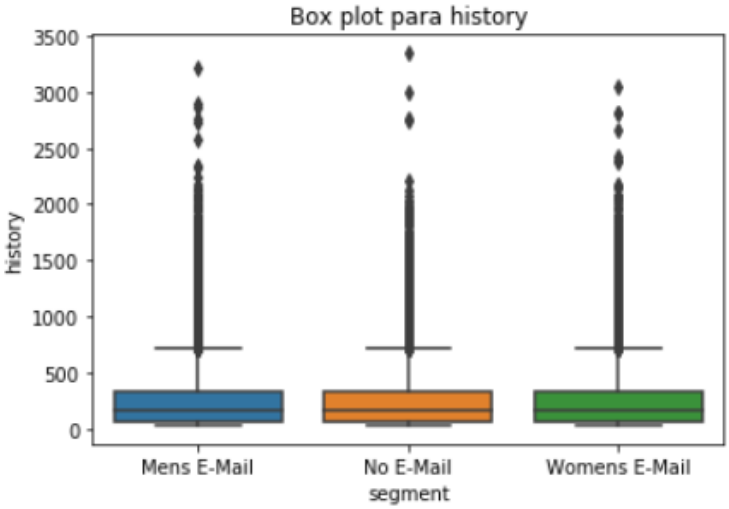
Tabla 3-3: Tabla de frecuencias relativas para ‘segment’ en 5 muestras aleatorias

Tratamiento	**Muestra 1**	**Muestra 2**	**Muestra 3**	**Muestra 4**	**Muestra 5**
Mens E-Mail	33.00%	33.80%	33.25%	33.21%	33.33%
No E-Mail	33.85%	33.48%	33.40%	32.98%	33.43%
Womens E-Mail	33.15%	32.72%	33.35%	33.81%	33.23%

También se puede evaluar si hay algún tipo de relación entre los valores de las variables y el recibir o no el tratamiento (alguno de los correos). En las tablas Tabla 6-1 a Tabla 6-8 en el Apéndice I se puede ver que el porcentaje de observaciones con cada tratamiento (control o cada una de las campañas) es similar para cada valor posible de las variables categóricas. El único atributo continuo que hay en los datos es 'history'. Una opción sería *binear* la variable y hacer un análisis similar a lo anterior. Pero no es necesario, ya que basta con mirar la tabla para la variable

'history_segment'. Como análisis adicional, se presenta un *boxplot* agrupando según el tratamiento. Las distribuciones de la variable continua para cada tratamiento parecen similares.

Figura 3-1: Box plot para 'history'



Otra manera de evaluar la independencia del tratamiento es hacer un test de hipótesis C2ST (*Classifier Two Sample Test*) [13], donde la hipótesis nula es que el *target* es independiente de los atributos. En el caso de la campaña de indumentaria femenina se obtuvo $p_W = 0.25$, y para la campaña de indumentaria masculina se obtuvo $p_M = 0.13$. Luego, en ambos casos no hay suficiente evidencia para rechazar Ho y se concluyó que se puede asumir que el *target* es independiente de los atributos.

En conclusión, en esta sección se presentó suficiente evidencia para poder asumir que el tratamiento es realmente aleatorio. Este resultado, además, permitirá asumir la validez de los supuestos sin necesidad de mayor análisis.

3.3. Análisis de supuestos

Como se comentó en la sección 2, al trabajar en el marco teórico propuesto por Rubin (1974) [18], se asumen varios supuestos listados en dicha sección. En el siguiente apartado, se analizó la razonabilidad de esos supuestos en los datos de Hillstrom.

El primer supuesto que se listó fue SUTVA, que esencialmente significaba que la probabilidad de que el individuo i pertenezca al grupo de tratamiento ($Ti = 1$) sólo depende de Di. Dado que los *potential outcomes* no se conocen en su totalidad, este supuesto no se puede validar completamente a partir de los datos, y se depende de la razonabilidad de suponer esto. En este caso, es lógico suponer que un individuo reciba o no algún correo electrónico de la campaña no va a depender del tratamiento o atributos de otro individuo. Una de las razones importantes para asumir SUTVA es que implica, por ejemplo, que dos individuos con los mismos atributos tienen la misma probabilidad de ser tratados, y ya se vio en el apartado 3.2 que es razonable asumir esto.

El segundo supuesto mencionado (Positividad u *Overlap*) es equivalente a asumir que para cada conjunto de atributos existen observaciones tanto en el grupo de control como en el de tratamiento. Cuantas más variables, más difícil es que se cumpla, más aún con variables continuas. En este caso, se cuenta con solamente 8 atributos, todas variables categóricas excepto una. Además, como ya se mostró evidencia de que el tratamiento es completamente aleatorio, y que no depende de las características, se puede asumir un *propensity score* constante. Esto es, se puede considerar que la probabilidad de ser tratado es constante (y no trivial), cumpliéndose el supuesto.

El último supuesto a considerar es el de Independencia Condicional (*Unconfoundedness*), el cual se va a cumplir en experimentos donde la asignación del tratamiento es aleatoria, como es en este caso.

3.4. A primera vista: ATE

En esta subsección se pretende generar una primera intuición sobre los resultados de las campañas. En resumen, la población total fue dividida aleatoriamente en tres grupos: los clientes que recibieron un correo electrónico de mercadería femenina, los que recibieron uno de mercadería masculina y por último, los que no recibieron correo. En este trabajo se consideraron las siguientes dos campañas:

Tabla 3-4: Tabla de frecuencias – muestras de control y de tratamiento

Campaña de mercadería femenina	Observaciones #
$T_W = 1$	21.387
$T_W = 0$	21.306
	Total: 42.693

Campaña de mercadería masculina	Observaciones #
$T_M = 1$	21.307
$T_M = 0$	21.306
	Total: 42.613

Para analizar la efectividad de las campañas hay tres efectos a considerar (*outcomes* de interés): impacto en la cantidad de visitas al sitio, en las compras realizadas y en la cantidad gastada en dichas compras. En base a los resultados a continuación (Tabla 3-5), se puede concluir:

1) Ambas campañas son exitosas en el sentido que estas tres medidas aumentaron (ATE > 0)

2) Las tres efectos indican consistentemente que la campaña de mercadería masculina es más exitosa.

- Un aumento de 4.52 puntos porcentuales (pp) con la campaña femenina vs 7.66 pp con la campaña masculina;
- Un aumento de 0.31 pp con la campaña femenina vs 0.68 pp con la campaña masculina;
- Un aumento de $0.42 por individuo con la campaña femenina vs $0.76 con la campaña masculina.

3) En términos de cambio porcentuales, la campaña masculina logra duplicar la cantidad de compras y el monto gastado, y aumentar en un 72% la cantidad de visitas al sitio.

Tabla 3-5: Impacto de las dos campañas

segment	visit	conversion	spend
Mens E-Mail	18.28%	1.25%	1.422617
No E-Mail	10.62%	0.57%	0.652789
Womens E-Mail	15.14%	0.88%	1.077202

segment	visit	conversion	spend
ATE T_W	4.52%	0.31%	0.424412
ATE T_M	7.66%	0.68%	0.769827

segment	visit	conversion	spend
Impacto T_W %	42.61%	54.33%	65.02%
Impacto T_M %	72.14%	118.84%	117.93%

Desde el punto de vista del negocio, un empresario puede no estar satisfecho ya que el aumento en las ganancias no es significativo. Sin embargo, la estimación del efecto promedio no brinda mucha más información ni sugiere como mejorar la(s) campaña(s). Para este tipo de toma de decisiones es que es más útil estimar el efecto individual o condicional a ciertas características de

los individuos. De esta manera se puede orientar futuras campañas a un público que se supone será más receptivo.

3.5. *Outcomes* de Interés

Las tablas a continuación muestran las frecuencias de visitas al sitio de internet y de compras realizadas posterior a la campaña. Como se ve, las variables dependientes tienen distribuciones bastantes desbalanceadas. La proporción de unos para 'visit' es 12.89% y 14.44% para la campaña femenina y masculina, respectivamente, lo cual es aceptable. Sin embargo, al mirar el número de compras, las proporciones son mucho menores, 0.73% y 0.91%, respectivamente. Más aún, si se mira la distribución de 'spend' en las observaciones con gasto positivo ('conversion=1'), los histogramas (Figura 3-2 y Figura 3-3), se observa una distribución muy asimétrica con gastos muy cercanos a $0. Esto va a significar problemas a la hora del modelado.

Tabla 3-6: Campaña Femenina – frecuencias para 'visit' y 'conversion'.

segment visit	No E-Mail	Womens E-Mail	All
0	19044	18149	37193
1	2262	3238	5500
All	21306	21387	42693

segment conversion	No E-Mail	Womens E-Mail	All
0	21184	21198	42382
1	122	189	311
All	21306	21387	42693

Figura 3-2: Distribución de 'spend' en los 311 clientes que compraron luego de la campaña de mercadería femenina

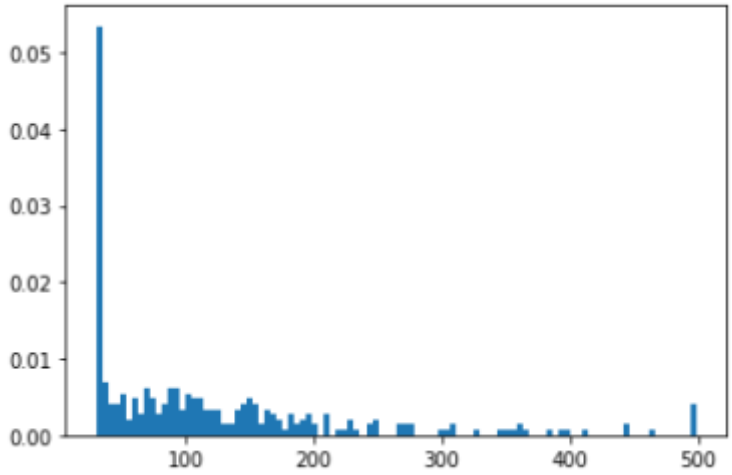
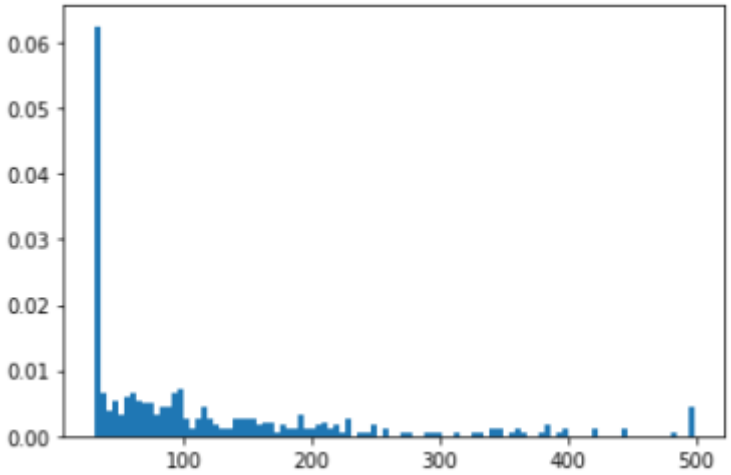


Tabla 3-7: Campaña Masculina – frecuencias para 'visit' y 'conversion'.

segment visit	Mens E-Mail	No E-Mail	All
0	17413	19044	36457
1	3894	2262	6156
All	21307	21306	42613

segment conversion	Mens E-Mail	No E-Mail	All
0	21040	21184	42224
1	267	122	389
All	21307	21306	42613

Figura 3-3: Distribución de 'spend' en los 389 clientes que compraron luego de la campaña de mercadería masculina



3.6. Muestreo

Antes de pasar al modelado en el capítulo siguiente, se definieron unas nuevas variables a partir de las existentes y se dividieron los datos (cada campaña) en una muestra de entrenamiento (80%) y en una de validación (20%), que se usó para todos los modelos considerados.

La lista final de variables se encuentra en la Tabla 6-9 en el Apéndice I. Esencialmente, se convirtieron las variables categóricas en factores, necesario para el modelado, como se verá más adelante. Además se definió el *propensity score* necesario para calcular el *Transformed Outcome,* fijándolo en 0.5 para cada observación dado que para cada campaña aproximadamente la mitad de los clientes recibieron el correo electrónico y el tratamiento fue completamente aleatorio.

La tabla a continuación muestra la cantidad de observaciones en cada muestra.

Tabla 3-8: Cantidad de observaciones por muestra. (Izquierda: campaña masculina, derecha: campaña femenina)

Muestra	#
MEN_CAMPAIGN	42613
MEN_train	34090
MEN_test	8523

Muestra	#
WOMEN_CAMPAIGN	42693
WOMEN_train	34154
WOMEN_test	8539

Dado que la muestra de testeo se va a usar para analizar y evaluar las métricas de *performance*, es importante que la composición sea similar y representativa de la población total. En otras palabras, cuando un modelo entrenado en una población se usa en otra muestra (como podría llegar a ser una campaña futura), uno asume que la población se mantuvo estable. Para eso, se hizo un análisis de VSI[5] (por sus siglas en inglés, "*Variable Stability Index*", también llamado "*Characteristic Analysis*") para evaluar la migración de la población entre distintos valores o rangos de valores de las siguientes variables predictoras: 'recency', 'mens', 'womens', 'newbie', 'zip_code', 'channel' y 'history_segment'. En ambas campañas, en general se obtuvieron valores inferiores a 0.1 (que es lo recomendable [3]). Los cálculos completos se muestran en las tablas Tabla 6-10 y Tabla 6-11 del Apéndice I.

[5] VSI = $\sum_i [(test\% - train\%) \cdot \ln(test\%/train\%)]$, donde i recorre los valores (o bins) del predictor.

4 Modelos y estimaciones

Las técnicas de modelos de efectos de tratamiento se pueden clasificar en dos grandes grupos: estimación directa o indirecta. Como se probó en el Teorema 1 en la sección 2, apoyándose en los supuestos de Rubin de *potential outcomes*, para $\mu_t(x) = \mathbb{E}[Y^{obs} \mid X = x,\ T = t]$ vale la siguiente igualdad:

$$CATE(x) = \mu_1(x) - \mu_0(x) \quad \forall\, x$$

Los métodos de estimación indirecta estiman por separado cada término minimizando el error cuadrático medio en cada caso, y luego restando las predicciones obtenidas:

$$\widehat{CATE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x) \quad \forall\, x$$

Por otro lado, los métodos de estimación directa se basan en minimizar la directamente la pérdida $\mathbb{E}\left[\left(CATE(x) - \widehat{CATE}(x)\right)^2\right]$.

Ambas perspectivas tienen sus falencias. En el primer caso, la estimación es ineficiente ya que no se está minimizando en el error de real interés, sino en dos errores auxiliares. Esto significa que, para que la estimación sea precisa, es necesario que ambas estimaciones intermedias lo sean; de lo contrario los errores pueden amplificarse en el modelo agregado. Además, Radcliffe y Surry [16] sostienen que estos métodos no funcionan bien en la práctica, ya que la estimación obtenida no es el resultado de haber modelado el *uplift* explícitamente, sino la variable *outcome* en cada grupo de tratamiento. De hecho, se pudo haber dejado afuera del modelo a variables directamente relacionadas con el *uplift*.

En el caso de la estimación directa del *uplift*, sí se pretende minimizar el error de interés, pero sin embargo éste parece ser inviable por el problema fundamental de la inferencia causal ya planteado. Sin embargo, ya se comentó en la sección 2.1.1 cómo se puede sortear este obstáculo.

En las siguientes subsecciones, se consideran esencialmente tres tipos de modelos, dos de los cuales son de estimación directa (KNN y Random Forests) y uno indirecta (regresiones lineales).

Como ya se mencionó en la sección 3.1, se trabajó por separado en las dos campañas distinguidas en los datos (campañas de mercadería femenina y masculina), aplicando los mismos métodos de modelado y criterios. De esta manera, el trabajo sirve como análisis de dos experimentos distintos, y en última instancia se podrá evaluar si hubo diferencias en las conclusiones en las dos campañas, y a qué se debe.

Además, para cada campaña, se entrenaron y compararon modelos para el *uplift* en los tres posibles *outcomes* de interés. Si bien las técnicas usadas son las mismas, las conclusiones varían.

4.1. Regresión Lineal

4.1.1 Regresión Lineal – Background Teórico

Una de las técnicas más tradicionales para estimar efectos de tratamientos es usar regresiones lineales.

A modo de introducción, considerar que se quiere estimar el ATE y se propone el siguiente modelo lineal:

$$Y^{obs} = \beta_0 + \beta_1 T + \varepsilon, \qquad \mathbb{E}(\varepsilon \mid T) = 0$$

Entonces, asumiendo *Unconfoundedness*,

$$\mathbb{E}(Y(0)) = \mathbb{E}(Y(0) \mid T = 0) = \beta_0$$

$$\mathbb{E}(Y(1)) = \mathbb{E}(Y(1) \mid T = 1) = \beta_0 + \beta_1$$

Luego,

$$ATE = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0)) = \beta_1$$

Si se supone ahora que se tiene una sola característica observable, X, se puede considerar el siguiente modelo lineal:

$$Y^{obs} = \beta_0 + \beta_1 T + \beta_2 X + \varepsilon, \qquad \mathbb{E}(\varepsilon|X,T) = 0$$

Entonces, nuevamente asumiendo *Unconfoundedness*,

$$\mathbb{E}(Y(0)|X = x) = \mathbb{E}(Y(0)|X = x, T = 0) = \beta_0 + \beta_2 x$$

$$\mathbb{E}(Y(1)|X = x) = \mathbb{E}(Y(1)|X = x, T = 1) = \beta_0 + \beta_1 + \beta_2 x$$

Luego,

$$CATE(x) = \mathbb{E}(Y(1)|X = x) - \mathbb{E}(Y(0)|X = x) = \beta_1$$

En ambos casos, se obtiene el efecto promedio de tratamiento estimando β_1 por mínimos cuadrados, y se puede testear estadísticamente la existencia de efecto de tratamiento (t-test para β_1). La segunda estimación intenta captar el impacto de la variable X en el efecto de tratamiento, pero de todas maneras devuelve una estimación constante para toda la población. Para estimar el efecto heterogéneo, hay que incorporar el término de interacción entre la variable de interés, es decir, aquella que puede explicar la heterogeneidad, y el indicador del tratamiento. Se propone entonces el siguiente modelo lineal alternativo:

$$Y^{obs} = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 T \cdot X + \varepsilon, \qquad \mathbb{E}(\varepsilon|X,T) = 0$$

Operando como antes, se obtiene que:

$$\mathbb{E}(Y(0)|X = x) = \beta_0 + \beta_2 x$$

$$\mathbb{E}(Y(1)|X = x) = \beta_0 + \beta_1 + (\beta_2 + \beta_3)x$$

Luego,

$$CATE(x) = \mathbb{E}(Y(1)|X = x) - \mathbb{E}(Y(0)|X = x) = \beta_1 + \beta_3 x$$

Este modelo logra captar el impacto de la variable X en el efecto de tratamiento, generando una estimación con tantos valores distintos como tenga X. Se concluye que el efecto de tratamiento depende de las variables donde el coeficiente del término de interacción es estadísticamente significativo.

Además, cabe resaltar que entonces el estimador del CATE hereda las propiedades que se conocen del estimador de MCO, como la consistencia.

Claro que en este análisis las conclusiones están limitadas a haber incluido una sola variable. Sin embargo, en general se cuenta con más de una variable de interés. Al incluir más variables en la regresión, también hay que incluir los correspondientes términos de interacción, entre las variables y el indicador, y entre las distintas variables, aumentando la cantidad de parámetros a estimar. Considerando dos variables, el modelo lineal quedaría de la siguiente manera, necesitando estimar 8 parámetros:

$$Y^{obs} = \beta_0 + \beta_1 T + \beta_2^1 X^1 + \beta_2^2 X^2 + \beta_3^1 T \cdot X^1 + \beta_3^2 T \cdot X^2 + \beta_4 T \cdot X^1 \cdot X^2 + \beta_5 X^1 \cdot X^2 + \varepsilon,$$
$$\mathbb{E}(\varepsilon|X,T) = 0$$

En general, si se tienen p variables (además del tratamiento), se tendría la siguiente cantidad de parámetros:

$$\sum_{k=0}^{p+1} \binom{p+1}{k}$$

Este método entonces tiene poco poder estadístico y puede sufrir problemas computacionales. Además, se impone una estructura a la forma funcional.

4.1.2 Regresión Lineal – Modelado

Para empezar y a modo de ejemplo, a continuación se muestran los resultados de la estimación por MCO del ATE sin características para cada *output* de interés. Los resultados en la Tabla 4-1 son sobre el total de la población para explicitar que los resultados obtenidos coinciden con los valores en la Tabla 3-5.

Tabla 4-1: Estimación del ATE por MCO sin características para ambas campañas (T_w= femenina, T_m = masculina)

T_w	ATE param	t stat	p val	R^2	R^2 adj
Visit	0.045233	13.980750	0.0000	0.004558	0.004534
Conversion	0.003111	3.780105	0.0002	0.000335	0.000311
Spend	0.424412	3.254768	0.0011	0.000248	0.000225

T_m	ATE param	t stat	p val	R^2	R^2 adj
Visit	0.076590	22.620112	0.0000	0.011865	0.011842
Conversion	0.006805	7.389672	0.0000	0.001280	0.001256
Spend	0.769827	5.300090	0.0000	0.000659	0.000635

Para obtener una estimación del CATE, para cada campaña y para cada *output* de interés, se estimó un modelo de MCO con sólo una característica (y su interacción), debido a la creciente complejidad de la forma funcional al agregar más atributos. Por este mismo motivo, sólo se consideraron predictores numéricos (es decir: continuos o *dummies*). Se estimó tanto en las muestras totales ('WOMEN_CAMPAIGN' y 'MEN_CAMPAIGN') como en las respectivas submuestras de entrenamiento, para evaluar la estabilidad de los coeficientes (ver Tabla 7-1 - Tabla 7-4 en el Apéndice II). Al comparar los resultados empiezan a hacerse evidente las falencias del método: se puede observar variabilidad en los coeficientes obtenidos y en la mayoría de los casos no se pasa el test de significatividad individual. En este sentido, los resultados para el *outcome* de interés 'visit' son los más satisfactorios.

Ya que no hay indicios claros en los resultados obtenidos de cuál es la especificación superadora, o siquiera de si la hay, para seleccionar las siguientes especificaciones además de mirar los p-valores, también se miraron las métricas de importancia de variables obtenidas con los Random Forests que se analizarán luego[6], y se compararon en base a MSE. La alternativa de quitar el término de interacción no permite sacar conclusiones sobre qué público se vio más influenciado por la campaña; por otro lado, la alternativa de agregar más atributos y sus respectivas interacciones sólo agrega complejidad a un método que no parece demasiado prometedor.

Tabla 4-2: Especificaciones elegidas (en base a los datos de entrenamiento)

Outcome de interés	Campaña femenina	Campaña masculina
Visit	Atributo elegido: mens $\beta_T = 0.071868$ $\beta_{XT} = -0.046177$	Atributo elegido: history $\beta_T = 0.071979$ $\beta_{XT} = 0.000025$
Conversion	Atributo elegido: mens $\beta_T = 0.004297$ $\beta_{XT} = -0.002397$	Atributo elegido: history $\beta_T = 0.005530$ $\beta_{XT} = 0.000006$
Spend	Atributo elegido: mens $\beta_T = 0.444164$ $\beta_{XT} = -0.188730$	Atributo elegido: history $\beta_T = 0.514608$ $\beta_{XT} = 0.000965$

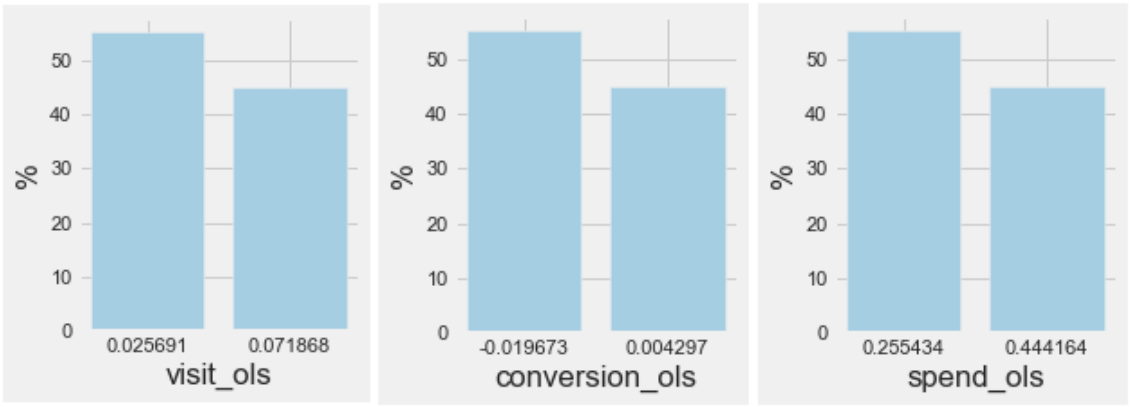
[6] Esencialmente, 'history' parece ser significativamente más importante que todas las otras variables para toda combinación de campañas y outcome, excepto en la predicción del uplift en 'visit' para la campaña femenina, que es 'womens' (ver Tabla 4-3 a Tabla 4-8, en la sección 4.3.3).

En todos los casos, el coeficiente del tratamiento (β_T) es positivo, lo cual es de esperarse: el tratamiento impactó en las medidas de interés. En el caso de la campaña de mercadería femenina, el atributo elegido para cada modelo fue 'mens' y siempre se obtuvo un coeficiente negativo en la interacción con el tratamiento (β_{XT}), lo que parece sugerir que la campaña tuvo más efecto en clientes que no compraron mercadería masculina, es decir, sólo compraron mercadería femenina en el último año. Para la otra campaña, el efecto del tratamiento se modeló con el atributo 'history' y para los tres *outcomes* se obtuvo un coeficiente positivo sugiriendo una relación directa entre el monto gastado en compras anteriores y el efecto de la campaña.

Para cada campaña, las predicciones obtenidas se llaman en base al *outcome* considerado: 'visit_ols', 'conversion_ols' y 'spend_ols'.

En los casos de la campaña femenina (Figura 4-1), como la variable elegida es una *dummy*, solo se obtienen dos predicciones posibles, que a grandes rasgos divide a la población en 55% - 45%. Por otro lado, para la campaña masculina, se utilizó para los tres *outcomes* a la variable 'history' como predictor, por lo que la predicción es, de alguna manera, continua.

Figura 4-1: Distribución de las predicciones por OLS – Campaña Femenina



Nota: Se verá más adelante que los grupos determinados por cada una de las estimaciones coinciden.

Más allá de que los estadísticos (ver Tabla 7-1 - Tabla 7-4 en el Apéndice II) de las especificaciones elegidas no son buenos, tampoco son útiles a la hora de comparar con distintas metodologías. En la sección 4.4 se compararán todos los métodos propuestos en base a las métricas definidas en la sección 2.1 y a la capacidad de ordenamiento de la población.

En las secciones a continuación se presentan ejemplos de métodos directos de estimación del CATE que permiten introducir una mayor cantidad de variables.

4.2. KNN

4.2.1 KNN – Background Teórico

Dada la imposibilidad de conocer los contra-factuales, una alternativa sería encontrar para cada unidad en el grupo de control, una unidad "espejo" en el grupo de tratamiento que tenga exactamente los mismos atributos para estimar el contra factual. Análogamente, se repite para cada individuo en el grupo de tratamiento. El supuesto de *Unconfoundedness* garantiza que los *outcomes* de dichos pares de individuos se pueden pensar como los dos *outcomes* pontenciales de un sólo individuo. Esta técnica se la conoce como "*Exact Matching*". Más en general y para evitar la discrecionalidad de elegir un solo individuo con idénticas características, para cada observación en el grupo de control (respectivamente, en el grupo de tratamiento), se puede promediar el *outcome* de todos sus "espejos" en el grupo de tratamiento (respectivamente, grupo de control) para predecir el contra factual.

Sin embargo, es poco probable que exista para cada observación en el grupo de control un *match* exacto en el grupo de tratamiento, y viceversa, especialmente cuando las características X son

continuas, o aumenta la cantidad de predictores. Es más razonable buscar observaciones con *matching* aproximado. Nuevamente, dada una observación *i* en el grupo de tratamiento, se puede buscar una observación j_i en el grupo de control de tal manera que $\|X_i - X_{j_i}\|$ sea mínimo, para alguna noción de distancia elegida, y de esa manera imputar el contra factual. Incluso se puede dar un paso más, en pos de generalizar, y en vez de elegir un único elemento cercano, nuevamente elegir varios e imputar el contra factual usando el promedio.

El método de regresión "Causal KNN", introducido por Hitsch y Misra (2018) [10] está representado con la siguiente ecuación

$$CATE_K\ (x)\ =\ \frac{1}{K}\sum_{i\in N_K(x,1)} Y_i\ -\ \frac{1}{K}\sum_{i\in N_K(x,0)} Y_i \tag{5}$$

Donde $N_K(x,t)$ denota al conjunto de los K vecinos más cercanos a un individuo de características $X = x$, con tratamiento $T = t$, e Y_i son los valores observados de la variable de interés. La estimación depende del valor del hiperparámetro K ≥ 1. El máximo valor posible está determinado por el tamaño del menor grupo de tratamiento. Un mayor valor de K aumenta la cantidad de observaciones consideradas para los promedios, disminuyendo la similitud con las características $X = x$. Va a ser importante elegir un valor apropiado para este hiperparámetro. En su paper Wager y Athey [19] también usan la técnica de KNN como *benchmark* para Causal Random Forests, considerando directamente los valores K = 10 y K = 100. En esta tesis, en cambio, se buscó un valor óptimo de K considerando como criterio de elección el error cuadrático medio a partir del *transformed outcome loss*, métrica explicada en la sección 2.1.1. En el apartado siguiente, se describe en más detalle el modelado con esta técnica.

4.2.2 KNN - Modelado

Para implementar esta técnica, se utilizó la función "KNeighborsRegressor" del paquete Sklearn de Python [21]. Como hace falta elegir un K óptimo, dado un rango de valores considerados, se creó una función en Python que busca para cada observación en los datos de testeo los vecinos más cercanos en el grupo de control y en el de tratamiento de los datos de entrenamiento. A partir de sus correspondientes valores de Y (para cada *outcome* de interés: 'visit', 'conversion' y 'spend'), la función calcula el $CATE_K(x)$ para distintos valores de K y devuelve los valores de MSE. Luego, el usuario elige el valor de K que minimiza el error, o a partir del cual no hay mejora considerable.

Idealmente, se elegiría K tal que la expresión (6) a continuación sea mínima, pero como el verdadero efecto de tratamiento no es observable, esto es imposible.

$$\mathbb{E}\left[\left(CATE(x)\ -\ CATE_K(x)\right)^2\right] \tag{6}$$

Para sortear este obstáculo, se puede recurrir a la igualdad (4) (Propiedad 2) que implica que minimizar en K el error cuadrático medio de (6) es equivalente a minimizar el TOL:

$$\mathbb{E}\left[\left(Y_i^* - CATE_K(X)\right)^2\right]$$

Vale la pena mencionar que Causal KNN puede no ser apropiado en casos donde los datos (proporción de muestra de control y de tratamiento) son extremadamente desproporcionados, ya que en esos casos se dificulta la búsqueda de vecinos cercanos en alguno de los grupos de tratamiento (tratamiento o control). En este caso de estudio, no se tiene esta dificultad ya que en ambas campañas los grupos de control y de tratamiento son aproximadamente del mismo tamaño (ver Tabla 3-4).

Para el modelado se utilizó la métrica Euclídea como noción de distancia para elegir los vecinos cercanos y se consideraron las siguientes variables[7]: 'recency', 'history', 'mens', 'womens', 'newbie', 'zip_code_Rural', 'zip_code_Surburban', zip_code_Urban', 'channel_Multichannel', 'channel_Phone' y 'channel_Web'. Si bien se cuenta con una cantidad limitada de variables predictoras, los resultados de Wager y Athey [19] muestran que un mayor número de predictores no minimiza el MSE.

Las figuras a continuación muestran el proceso de elección del K óptimo en cada caso (para cada campaña, y cada *outcome*), en base al MSE (o TOL), el cual incluyó una instancia de refinamiento del rango considerado para K (figuras de la derecha). En cada caso se eligió un valor K (especificados en los títulos de las figuras) donde el error en la muestra de testeo alcanza un mínimo, o a partir del cual no se observan mejoras considerables o la curva del error ya no se comporta de manera monotónicamente decreciente.

Para las dos campañas en general se observa que el error para 'visit' es cercano a 0.5 y el error para 'conversion' es cercano a '0.03' mientras que para 'spend' el error es demasiado grande. La diferencia en los errores se debe a lo balanceado/simétrico de la distribución de la variable *outcome* de interés. Esto se discutirá en más detalle más adelante en la sección 4.4, cuando se comparen todos los modelos considerados, junto con las métricas de performance y otros análisis.

Figura 4-2: Elección de K óptimo (K = 275) – Campaña Femenina – 'Visit'

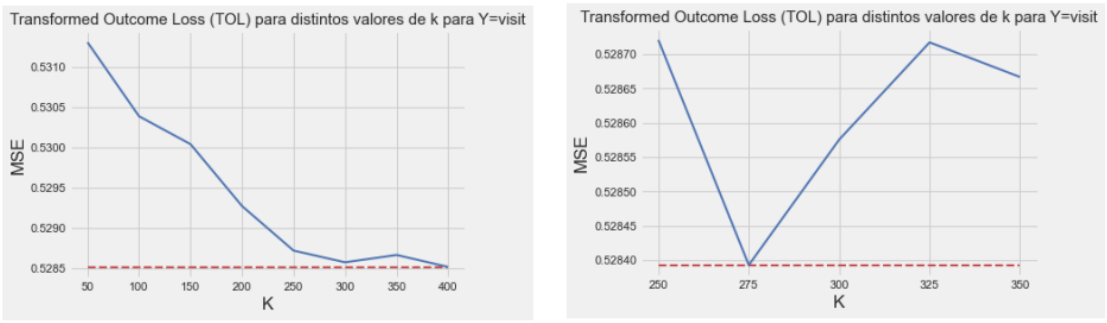
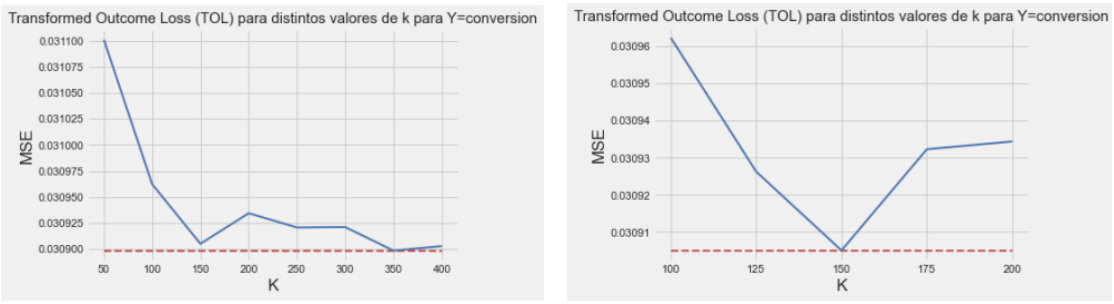


Figura 4-3: Elección de K óptimo (K = 150) – Campaña Femenina – 'Conversion'



[7] Todas las variables deben ser numéricas, para poder aplicar la distancia entre observaciones.

Figura 4-4: Elección de K óptimo (K = 150) – Campaña Femenina – 'Spend'

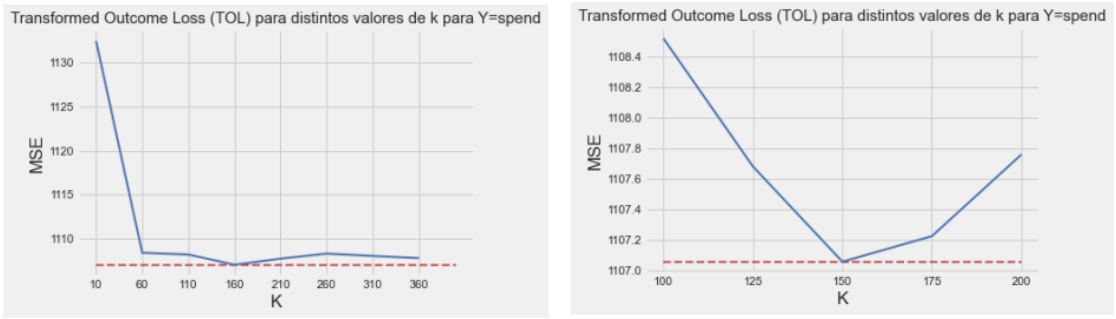


Figura 4-5: Elección de K óptimo (K = 275) – Campaña Masculina – 'Visit'

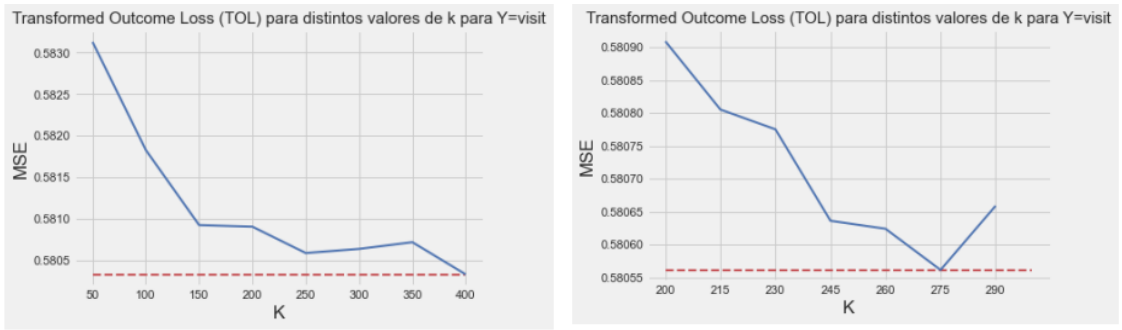


Figura 4-6: Elección de K óptimo (K = 200) – Campaña Masculina – 'Conversion'

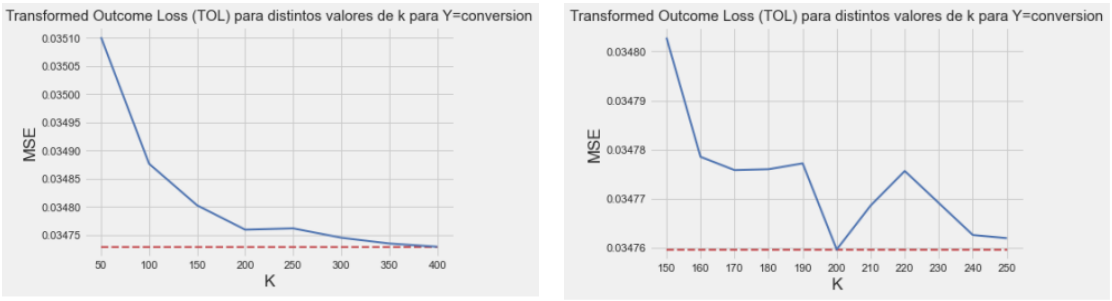
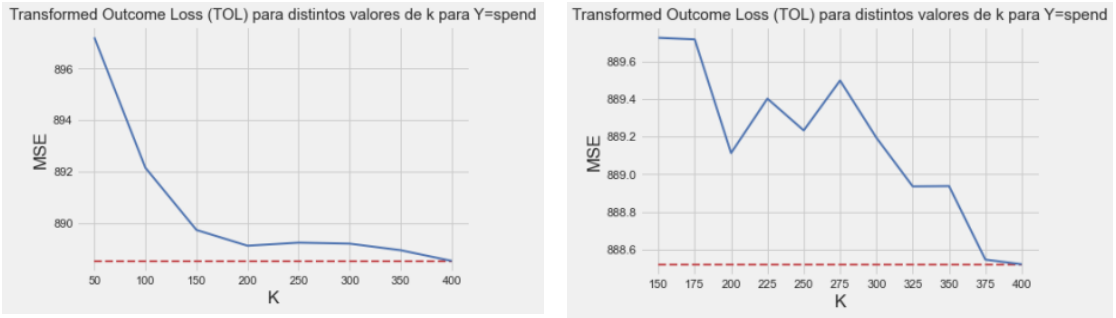


Figura 4-7: Elección de K óptimo (K =200) – Campaña Masculina – 'Spend



Para cada campaña, las predicciones obtenidas se llaman en base al *outcome* considerado y el K elegido, por ejemplo la estimación del CATE en 'spend' en la campaña masculina se llama 'spend_knn_200'.

4.3. Causal Random Forest

4.3.1 Honest Causal Random Forest – Background Teórico

El modelo que se desarrolla a continuación se puede pensar como un método de vecinos cercanos adaptativo, donde los datos determinan qué dimensiones son más importantes considerar al elegir los vecinos más cercanos. En KNN, la noción de cercanía estaba asociada a una métrica determinada (ej.: Euclídea fue la elegida en este caso). En los árboles de decisión, dos observaciones están cerca si caen en la misma hoja. Si bien las hojas están determinadas por los atributos y la distancia Euclídea, su tamaño y dirección de crecimiento está influenciado por la importancia de las características.

En la literatura hay resultados de varios autores usando Random Forests para estimar efectos de tratamiento. Por ejemplo, Foster y Ruberg (2011) [8] los utilizan para estimar el efecto de los atributos en el *outcome* de interés en el grupo de tratamiento y de control por separado, y luego restando, infieren el efecto de tratamiento para cada individuo. Es decir, utilizan los árboles como método de estimación indirecta, en el sentido desarrollado al comienzo de la sección 4. En cambio, Athey e Imbens (2016) [2] modificaron el algoritmo estándar de Random Forest, para estimar el efecto de tratamiento de manera directa. Estos enfoques no contaban con resultados formales en cuanto a inferencia estadística. En cambio, el nuevo enfoque de Wager y Athey [19] incorpora una cualidad a los árboles (llamada Honestidad) que junto con los supuestos del marco teórico de los *potential outcomes* propuesto por Rubin, como *Overlap* y *Unconfoundedness*, permite obtener estimaciones consistentes, insesgadas y asintóticamente normales.

Mientras que los árboles de decisión clásicos determinan los cortes en base a métricas como el error cuadrático, el índice de Gini o entropía, en los árboles causales que se consideraron en este trabajo los cortes se eligen para maximizar la variación del *uplift*[8] entre las hojas y aun así, generar estimaciones precisas de los efectos de tratamiento. Es más conveniente construir un Random Forest ensamblando varios árboles más sencillos en comparación de un sólo árbol más complejo y optimizado [5]. Una vez determinado el criterio de ramificación, los bosques se construyen de la manera clásica:

- Dentro de un árbol, para cada hoja L, y cada $x \in L$, se estima al efecto de tratamiento de la siguiente manera:

$$\widehat{CATE}(x) = \frac{1}{\#\{i: T_i = 1,\ i \in L\}} \sum_{\{i:T_i=1, i\in L\}} Y_i - \frac{1}{\#\{i: T_i = 0,\ i \in L\}} \sum_{\{i:T_i=0, i\in L\}} Y_i \qquad (7)$$

- La estimación generada por un Causal Random Forest conformado por B árboles considerando distintos subconjuntos de variables dependientes, es el promedio de las estimaciones de los árboles que lo conforman:

$$\widehat{CATE}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{CATE}_b(x)$$

El gran aporte de Wager y Athey [19] en su *paper* se debe a la construcción de árboles "honestos".

Definición 2: Se dice que un árbol es honesto si el valor del *outcome* (Y) de cada observación en la muestra de entrenamiento es utilizada o bien para determinar un corte o bien para calcular la estimación del efecto de tratamiento dentro de su hoja, pero no para ambas instancias.

Esta cualidad, permite reducir el sesgo y *overfitting*. Los autores describen dos ejemplos de árboles con esta cualidad. El árbol "Propensity" y el "double-sample".

En el primer tipo de árbol, se ignora el valor de Y al momento de decidir las ramificaciones y se entrena un árbol de clasificación para el tratamiento *T,* y el *outcome* sólo se utiliza a la hora de

[8] Para más detalle, referirse a la explicación de los árboles "double-sample".

calcular la estimación en cada hoja. En este tipo de árboles honestos, como está construido como árbol de clasificación, los cortes de deciden de manera clásica, por ejemplo utilizando el índice de Gini. Los autores explican que este enfoque es ideal cuando se quiere minimizar el sesgo causado por variaciones en el *propensity score* $e(x)$. Como esta no es la situación en los datos de Hillstrom, donde el *propensity score* se puede considerar constante sin pérdida de generalidad, en esta tesis se utilizó la segunda variante de árboles causales, explicada a continuación.

En el caso de los árboles con doble muestreo ("double-sample"), la muestra de entrenamiento se divide aleatoriamente en dos mitades, una para el proceso de *splitting* (que se notará S) y la otra para la estimación dentro de cada hoja (E). De esta manera, cada observación es solamente usada en una de las dos instancias de creación de un árbol. Una vez determinada la estructura del árbol, en cada nodo final, se calcula la estimación del CATE utilizando la fórmula (7) en las observaciones de la muestra E. Cabe resaltar que al momento de construir un Random Forest, para cada árbol, cada observación va a pertenecer aleatoriamente a una u otra muestra S o E, siendo utilizada entonces tanto para determinar los cortes como para la estimación, de manera agregada. De esta manera, no hay problemas de desperdicio de datos por el muestreo. En este tipo de árboles honestos los cortes se deciden de tal manera de maximizar la varianza de las estimaciones, inspirado en el hecho de que para árboles de regresión encontrar el corte que minimice el ECM es equivalente a que maximice la suma de los cuadrados de la estimación (que es el promedio, en el caso de un árbol clásico, no causal). La demostración de este hecho se encuentra en el Apéndice II -1.

4.3.2 Honest Causal Random Forest – DML

DML ("Double Machine Learning") [6] es un método para estimar efectos de tratamiento basado en la residualización de la variable de tratamiento y la variable *outcome*. El método recibe su nombre dado que el problema de estimación se descompone en dos etapas predictivas: estimando modelos primarios y auxiliares.

El método asume las siguientes ecuaciones estructurales:

$$Y = \Theta(X) \cdot T + g(X) + \epsilon \ , \qquad \mathbb{E}[\epsilon \mid X] = 0$$

$$T = f(X) + \eta \ , \qquad \mathbb{E}[\eta \mid X] = 0$$

Restando $\mathbb{E}[Y \mid X] = \Theta(X) \cdot \mathbb{E}[T \mid X\,] + \ g(X)$ de la primera ecuación se obtiene:

$$Y - \ \mathbb{E}[Y \mid X] \ = \ \Theta(X) \cdot (T - \mathbb{E}[T \mid X\,]) + \epsilon$$

Entonces, se estiman los residuos:

1. $Y - \ \mathbb{E}[Y \mid X]$ (predicción del residuo del *outcome*)
2. $(T - \mathbb{E}[T \mid X\,]) = \eta$ (predicción del residuo del tratamiento)

En la siguiente etapa, se estima el efecto de tratamiento a partir de los residuos anteriores, utilizando, por ejemplo, la técnica "Honest Causal Random Forest".

4.3.3 Causal Random Forest – Modelado

Para todos los modelos de esta sección se trabajó con la librería EconML de Python [21], en particular con las funciones "grf.CausalForest"[9] y "dml.CausalForestDML", las cuales cuentan con la posibilidad de modelar árboles honestos. A estos modelos se los notará como HCRF y DML HF, respectivamente.

[9] Esta función es la implementación en Python de la función "causal_forest" del paquete "grf" de R de los autores Wager y Athey [19].

Además, a los efectos de comparación, también se contrastan los resultados con Random Forests para efectos causales, pero sin la cualidad de honestidad, mutando el parámetro correspondiente en la función "grf.CausalForest". A esta especificación se la llamará CRF.

Las funciones tienen varios hiperparámetros que optimizar y determinar. En el proceso de optimización, se compararon las *performances* de los modelos con el ECM basado en TOL. A continuación se detallan los más relevantes:

- Número de árboles (n_estimators): hiperparámetro a optimizar. Se fue ajustando la grilla de búsqueda en base a evidencia de *overfitting*[10].
- Número mínimo de observaciones por nodo (min_sample_leaf): hiperparámetro a optimizar. Se consideraron valores cercanos al propuesto por Wager y Athey (i.e. 5).
- Métrica de decisión para los cortes: Se eligió la opción 'het' para ser consistente con el criterio usado por Wager y Athey en su paper.
- Honestidad (honest): Se consideró ambas opciones (True y False), a los efectos de comparación.

Se consideraron las siguientes variables, todas numéricas: 'recency', 'history', 'mens', 'womens', 'newbie', 'zip_code_Rural', 'zip_code_Surburban', zip_code_Urban', 'channel_Multichannel', 'channel_Phone' y 'channel_Web'. Dado que se cuenta con una cantidad limitada de variables predictoras, no se impuso restricción sobre la cantidad mínima de variables a utilizar.

Una de las ventajas de estas funciones es que cuentan con una métrica de importancia de las variables, basada en la heterogeneidad que crean. Esencialmente, dicho criterio de importancia aumenta cada vez que una variable es usada para generar una nueva ramificación. Además, esta métrica está ponderada por la profundidad del corte.

Para cada caso (para cada campaña, y cada *outcome*) y para cada tipo de modelo considerado (CRF, HCRF y DML HF), esencialmente se siguieron los siguientes pasos:

1) Optimización de hiperparámetros (eventualmente, refinando la grilla)
2) Pre-elección del modelo mirando MSE tanto en la muestra de entrenamiento como en la testeo, y cálculo de la importancia de las variables (VI)
3) Consideración de un modelo más parsimonioso en base a VI.
4) Elección del modelo final.

Al mirar el MSE, se buscó elegir hiperparámetros de tal manera el error en la muestra de entrenamiento y de testeo alcance un mínimo, o a partir del cual no se observan mejoras considerables o las curvas del error ya no se comportan de manera monotónicamente decreciente. Dado que en general el error en ambas muestras no tiene el mismo comportamiento, también se tuvo en cuenta evitar el *overfitting*: es decir, que el error en la muestra de entrenamiento no se reduzca mientras no se observe mejoras en la muestra de testeo.

Para cada campaña, las predicciones obtenidas se llaman en base al *outcome* y especificación de modelo considerados, por ejemplo: 'visit_hcrf', 'visit_dml_hf' y 'visit_crf'.

A continuación se describe la elección de especificaciones para los tres tipos de Causal Random Forests considerados.

Campaña Femenina – visit

La figura a continuación muestra las curvas del error (MSE o TOL) en las muestras de entrenamiento (azul) y de testeo (roja) para distintos valores de la cantidad de árboles considerada en el HCRF. Mientras que el error en la muestra de entrenamiento decrece

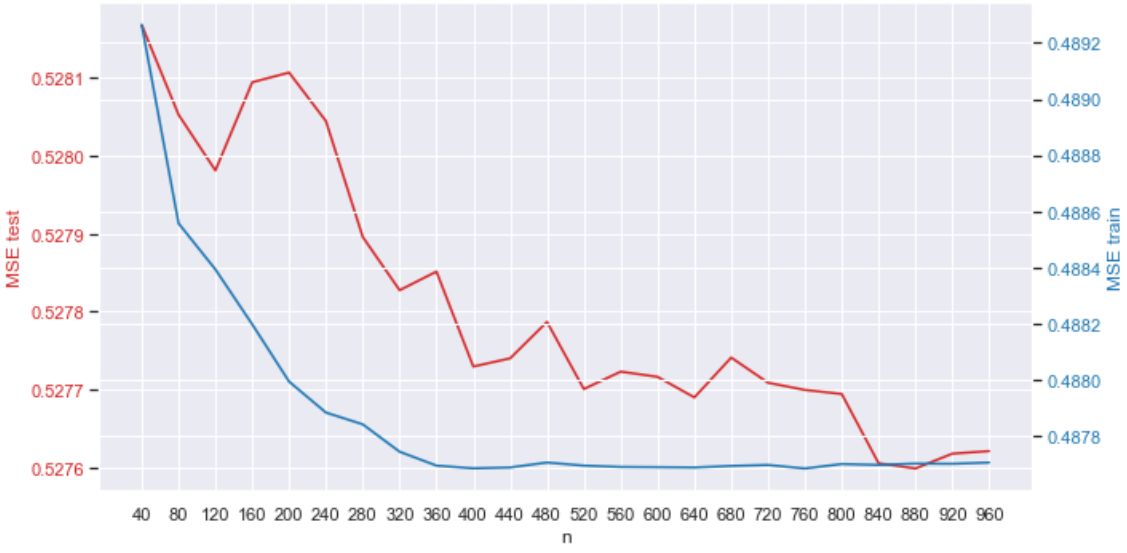
[10] Los gráficos en las secciones siguientes son ilustrativos y no cubren todos los rangos considerados.

constantemente, en la muestra de testeo el comportamiento es más errático. Se eligió n=440 ya que después de ese valor dicha curva oscila entre 0.5267 y 0.5268.

Figura 4-8: Elección hiperparámetros (min_samples_leaf=10, n_estimators=440) – Campaña Femenina – 'visit_hcrf'

En el caso del modelo DML HF, se eligió n=400 ya que a partir de ese valor el error en la muestra de entrenamiento no muestra mejoría y la velocidad de decrecimiento en la muestra de testeo es mucho menor, oscilando entre 0.5276 y 0.5278.

Figura 4-9: Elección hiperparámetros (min_samples_leaf=5, n_estimators=400) – Campaña Femenina – 'visit_dml_hf'



Los valores de MSE se mueven en rangos acotados y razonables. De todas maneras, se discutirá más en detalle esta métrica junto con otras a la hora de comparar todos los modelos en la sección 4.4.

La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. La documentación de EconML no tiene información sobre umbrales recomendados a la hora de elegir variables. De todas maneras, en esta y en las variantes siguientes (según campaña y *outcome*) se consideraron modelos más parsimoniosos excluyendo las variables peores rankeadas. Se observó que los MSE empeoraban, y dado que la cantidad disponible de variables es limitada, se optó por incluir todas las variables disponibles (las listadas al comienzo de la sección 4.3.3). Los tres modelos identifican a 'womens', 'history', 'recency' y 'mens', con alguna

variante en el orden, como las cuatro variables más importantes a la hora de predecir el *uplift* en la variable 'visit'. Además, se observa que 'womens' presenta un VI mucho mayor que la siguiente variable en cada caso.

Tabla 4-3: Importancia de las variables – Campaña Femenina – 'visit'

CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
Womens	0.68712	womens	0.644966	womens	0.668052
History	0.184457	mens	0.139504	History	0.133868
Recency	0.041863	history	0.13287	Mens	0.112022
Mens	0.03871	recency	0.036729	recency	0.042114
zip_code_Rural	0.014217	zip_code_Rural	0.012913	zip_code_Rural	0.010952
newbie	0.008031	zip_code_ Surburban	0.007776	zip_code_ Surburban	0.008137
zip_code_ Surburban	0.006845	newbie	0.00604	channel_Web	0.006324
channel_ Multichannel	0.006374	channel_Web	0.005967	newbie	0.005717
channel_Web	0.005271	channel_ Multichannel	0.004841	channel_ Multichannel	0.004669
channel_Phone	0.003749	channel_Phone	0.004522	zip_code_Urban	0.004077
zip_code_ Urban	0.003362	zip_code_Urban	0.003873	channel_Phone	0.004068

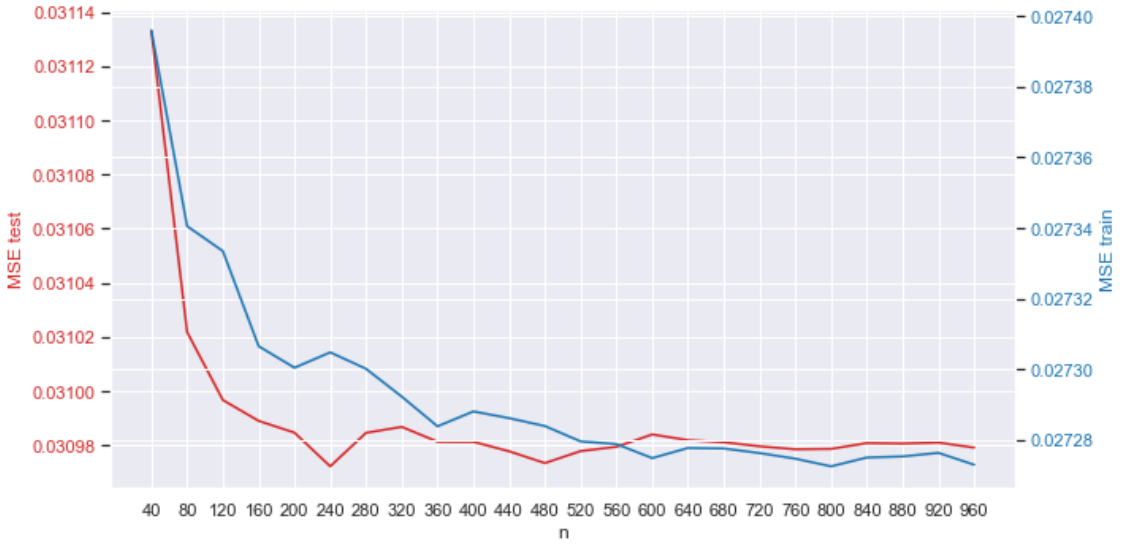
Campaña Femenina – conversion

En la Figura 4-10 se ve que, mientras que el error del modelo HCRF en la muestra de entrenamiento decrece en general, en la muestra de testeo el comportamiento se puede identificar un mínimo en n=240.

Figura 4-10: Elección hiperparámetros (min_samples_leaf=10, n_estimators=240) – Campaña Femenina – 'conversion_hcrf'

En el caso de DML HF (Figura 4-11) se eligió n=480 ya que a partir de ese valor el error en la muestra de testeo parece estabilizarse por encima, mientras que el error en la muestra de entrenamiento sigue decreciendo.

Figura 4-11: Elección hiperparámetros (min_samples_leaf=5, n_estimators=480) – Campaña Femenina – 'conversion_ dml_hf'



Los valores de MSE parecen sugerir que sin importar la elección de hiperparámetros, se está en la presencia de un buen modelo. Esto en general ocurrirá con todos los modelos (no sólo para RF) para el caso de 'conversion', como se discutirá más en detalle junto con otras métricas a la hora de comparar los modelos en la sección 4.4.

La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. Los tres modelos identifican de manera consistente a 'history', 'recency' y 'womens' como las tres variables más importantes a la hora de predecir el *uplift* en la variable 'conversion'. Además, se observa que 'history' presenta un VI mucho mayor que la siguiente variable, 'recency'.

Tabla 4-4: Importancia de las variables – Campaña Femenina – 'conversion'

CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
history	0.663743	history	0.575076	history	0.608456
recency	0.135626	recency	0.160273	recency	0.152326
womens	0.053696	womens	0.061983	womens	0.051665
channel_Phone	0.030036	zip_code_Rural	0.032361	channel_Phone	0.029347
mens	0.025508	channel_Phone	0.029976	zip_code_Rural	0.027362
zip_code_Rural	0.018533	channel_ Multichannel	0.026802	zip_code_Urban	0.026459
zip_code_Urban	0.016016	zip_code_Urban	0.025497	newbie	0.023395
newbie	0.015188	mens	0.025072	channel_ Multichannel	0.022251
zip_code_ Surburban	0.014635	newbie	0.023075	channel_Web	0.019676
channel_ Multichannel	0.014525	zip_code_ Surburban	0.022754	zip_code_ Surburban	0.019624
channel_Web	0.012494	channel_Web	0.017131	mens	0.019438

Campaña Femenina – spend

A diferencia de lo observado para modelos de 'visit' y 'conversion', en este caso los errores son demasiado importantes, como se puede ver en las figuras siguientes. Las causas, relacionadas con los datos, son análogas a aquellas por las que los errores de 'conversion' son muy pequeños, como se verá más adelante.

Figura 4-12: Elección hiperparámetros (min_samples_leaf=10, n_estimators=320) – Campaña Femenina – 'spend_hcrf'

Figura 4-13: Elección hiperparámetros (min_samples_leaf=5, n_estimators=280) – Campaña Femenina – 'spend_ dml_hf'

La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. Los tres modelos identifican de manera consistente a 'history', 'recency' y 'newbie' como las tres variables más importantes a la hora de predecir el *uplift* en la variable 'spend'. Además, se observa que 'history' presenta un VI mucho mayor que la siguiente variable, 'recency'.

Tabla 4-5: Importancia de las variables – Campaña Femenina – 'spend'

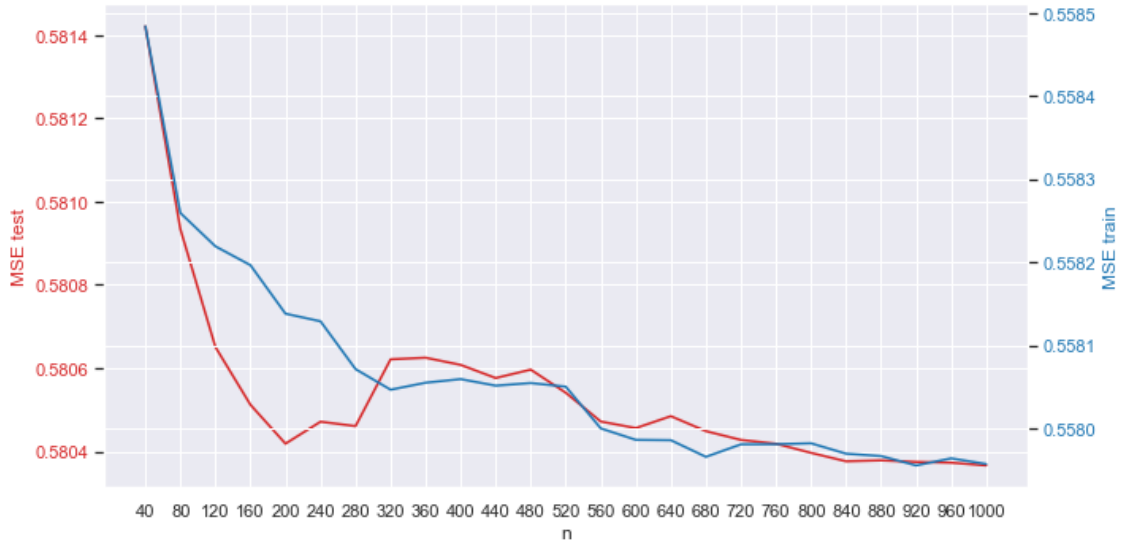
CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
History	0.645496	history	0.546113	History	0.543646
Recency	0.152351	recency	0.164522	Recency	0.179253
Newbie	0.046774	newbie	0.048455	Newbie	0.051264
womens	0.037992	channel_Phone	0.045612	zip_code_Urban	0.04576
channel_Phone	0.034219	womens	0.045463	womens	0.042127
zip_code_Urban	0.026404	zip_code_Urban	0.044268	channel_Phone	0.04203
zip_code_ Surburban	0.01292	zip_code_ Surburban	0.025124	zip_code_Rural	0.021069
Mens	0.012175	zip_code_Rural	0.02202	channel_ Multichannel	0.021022
channel_ Multichannel	0.010962	channel_ Multichannel	0.021355	zip_code_ Surburban	0.020575
channel_Web	0.010649	channel_Web	0.019262	Mens	0.019238
zip_code_Rural	0.01006	mens	0.017805	channel_Web	0.014015

Campaña Masculina – visit

En los modelos para la campaña masculina se operó de la misma manera.

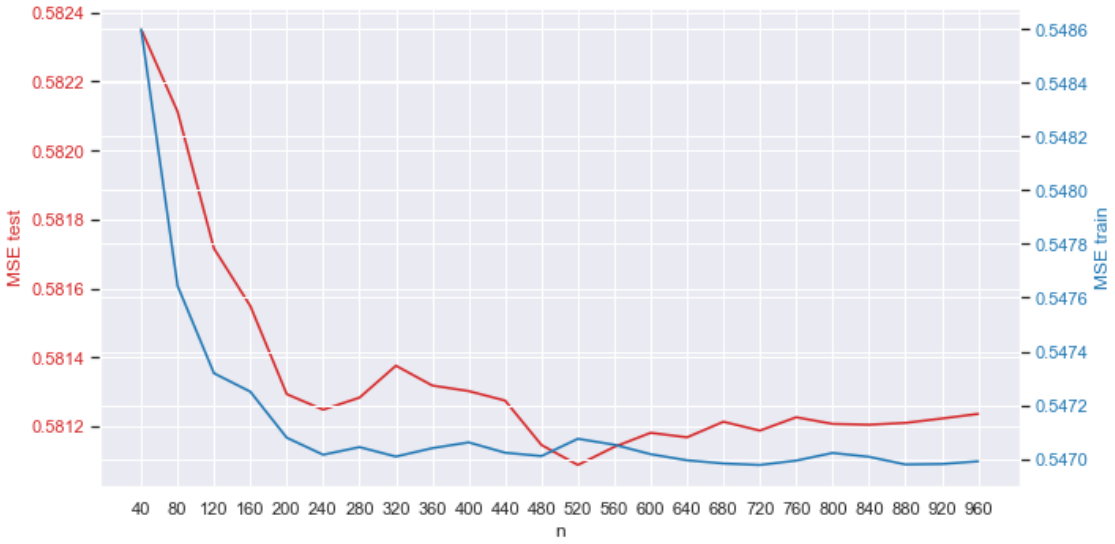
En el gráfico siguiente se muestran las curvas de MSE para HCRF; se eligió n=1000 ya que alrededor y luego (no visible en el gráfico) de dicho valor ambas curvas se estabilizan cerca de 0.5804 (test) y 0.587 (train).

Figura 4-14: Elección hiperparámetros (min_samples_leaf=10, n_estimators=1000) – Campaña Masculina – 'visit_hcrf'



En el caso de DML HF se eligió n=520 ya que a partir de ese valor el error en la muestra de testeo parece estabilizarse por encima.

Figura 4-15: Elección hiperparámetros (min_samples_leaf=5, n_estimators=520) – Campaña Masculina – 'visit_ dml_hf'



La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. Los tres modelos identifican de manera consistente a 'history', 'recency', 'womens', 'zip_code_Rural' y 'mens' como las cinco variables más importantes a la hora de predecir el *uplift* en la variable 'visit'. Además, se observa que 'history' presenta un VI mucho mayor que la siguiente variable, 'recency', y mucho mayor que 'womens', que fue identificado como la variable más importante en la campaña femenina.

Tabla 4-6: Importancia de las variables – Campaña Masculina – 'visit'

CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
History	0.655037	history	0.57991	history	0.583571
Recency	0.163294	recency	0.172007	recency	0.172193
womens	0.056348	womens	0.049675	womens	0.042335
zip_code_Rural	0.026607	zip_code_Rural	0.038378	zip_code_Rural	0.038529
Mens	0.024121	mens	0.034165	mens	0.031954
Newbie	0.018238	newbie	0.027064	zip_code_ Surburban	0.024588
channel_Phone	0.012017	channel_Web	0.02056	newbie	0.023384
channel_Web	0.011404	zip_code_Urban	0.020005	zip_code_Urban	0.021434
channel_ Multichannel	0.011232	channel_Phone	0.01972	channel_Phone	0.021411
zip_code_Urban	0.010949	zip_code_ Surburban	0.019622	channel_Web	0.02052
zip_code_ Surburban	0.010754	channel_ Multichannel	0.018894	channel_ Multichannel	0.020079

Campaña Masculina – conversion

Los gráficos siguientes muestran las curvas de MSE para HCRF y DML HF. Se pueden observar rangos de error similares a los presentes en el caso de la campaña femenina.

Figura 4-16: Elección hiperparámetros (min_samples_leaf=10, n_estimators=400) – Campaña Masculina – 'conversion_hcrf'

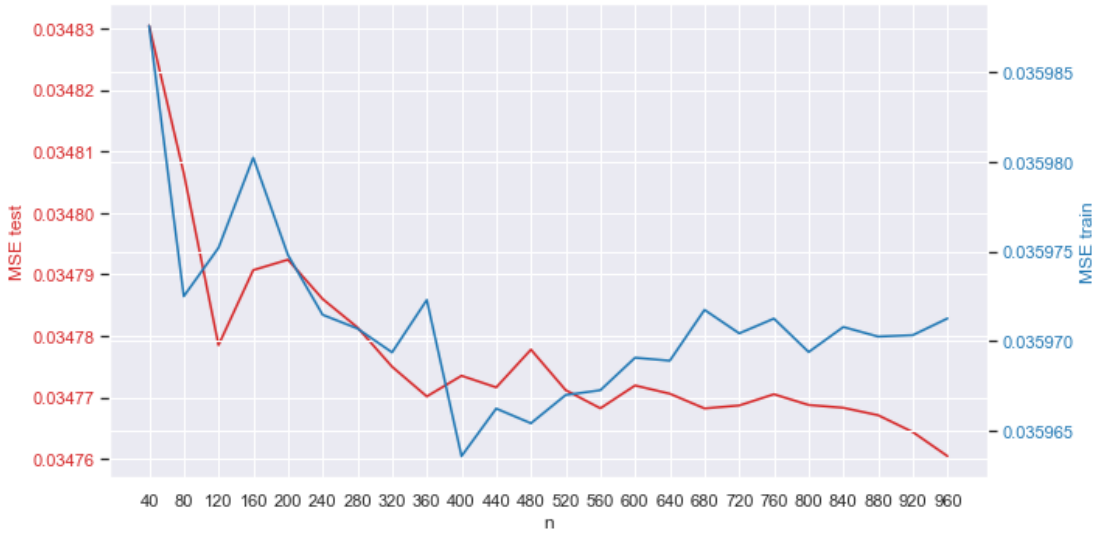


Figura 4-17: Elección hiperparámetros (min_samples_leaf=5, n_estimators=120) – Campaña Masculina a – 'conversion_ dml_hf'

La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. Los tres modelos identifican de manera consistente a 'history', 'recency' y 'zip_code_Rural' como las tres variables más importantes a la hora de predecir el *uplift* en la variable 'conversion'. Estas variables son similares a las observadas en el caso de la campaña femenina. También se observa que 'history' presenta un VI mucho mayor que la siguiente variable, 'recency'.

Tabla 4-7: Importancia de las variables – Campaña Masculina – 'conversion'

CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
history	0.682649	history	0.582397	History	0.566928
recency	0.106458	recency	0.15608	Recency	0.162354
zip_code_Rural	0.043192	zip_code_Rural	0.041084	zip_code_Rural	0.05453
womens	0.036122	newbie	0.039263	Newbie	0.04676
mens	0.028541	mens	0.032473	channel_Phone	0.034285
channel_Phone	0.027283	womens	0.03154	zip_code_Urban	0.029891
channel_Web	0.020953	channel_Phone	0.029649	Womens	0.028811
newbie	0.018943	zip_code_Urban	0.023764	Mens	0.022079
zip_code_Urban	0.015024	zip_code_ Surburban	0.023317	zip_code_ Surburban	0.019584
zip_code_ Surburban	0.013453	channel_Web	0.023067	channel_ Multichannel	0.018704
channel_ Multichannel	0.007381	channel_ Multichannel	0.017366	channel_Web	0.016074

Campaña Masculina – spend

Los gráficos siguientes muestran las curvas de MSE para HCRF y DML HF. Al igual que en el caso de la campaña femenina, se observan valores de error altos.

Figura 4-18: Elección hiperparámetros (min_samples_leaf=10, n_estimators=40) – Campaña Masculina – 'spend_hcrf'

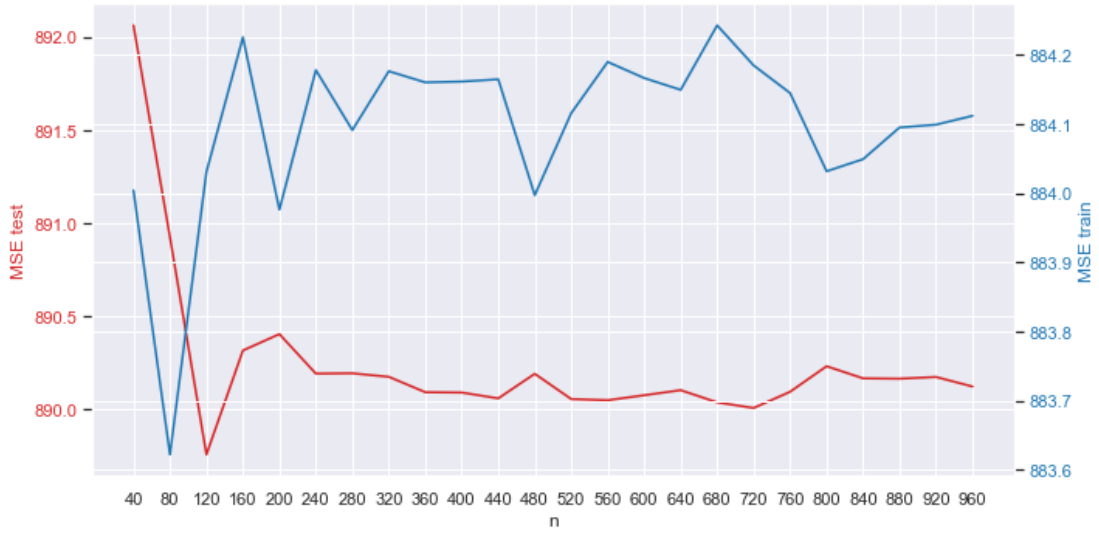
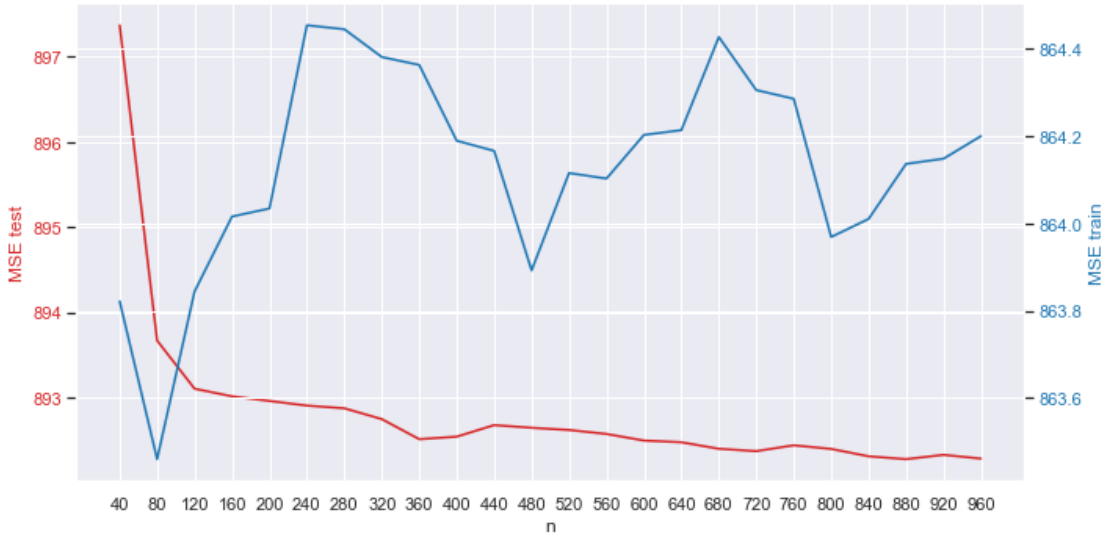


Figura 4-19: Elección hiperparámetros (min_samples_leaf=5, n_estimators=80) – Campaña Masculina – 'spend_dml_hf'



La tabla a continuación muestra los valores de VI para los tres modelos de RF considerados. Los tres modelos identifican de manera consistente a 'history' y 'recency' como las dos variables más importantes a la hora de predecir el *uplift* en la variable 'spend'. Estas variables son similares a las observadas en el caso de la campaña femenina. También se observa que 'history' presenta un VI mucho mayor que la siguiente variable, 'recency'.

Tabla 4-8: Importancia de las variables – Campaña Masculina – 'spend'

CRF		HCRF		DML HF	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
history	0.647467	history	0.516607	history	0.560492
recency	0.117464	recency	0.188484	recency	0.177867
womens	0.053028	channel_Phone	0.057456	channel_Phone	0.052332
channel_Phone	0.040487	newbie	0.053408	newbie	0.048673
mens	0.038098	womens	0.0443	womens	0.038723
newbie	0.035204	zip_code_ Surburban	0.031963	mens	0.031985
zip_code_ Surburban	0.023954	mens	0.030863	zip_code_Rural	0.023042
channel_Web	0.014989	channel_Web	0.027937	zip_code_ Surburban	0.022266
zip_code_Rural	0.010979	zip_code_Rural	0.023068	channel_Web	0.021156
channel_ Multichannel	0.010607	zip_code_Urban	0.017147	zip_code_Urban	0.016392
zip_code_Urban	0.007723	channel_ Multichannel	0.008768	channel_ Multichannel	0.007072

4.4. Comparación de Modelos

En esta sección se comparan todos los modelos considerados en base a las métricas de performance presentadas en la sección 2.1. Además, se definieron funciones ad-hoc a este trabajo como:

- "pred_CATE_by_bin": Para cada valor/rango de valores de las variables predictoras, se mira la estimación promedio del CATE en el *outcome* de interés. Así se obtiene una idea

de qué características están relacionadas con un mayor efecto de la campaña y en qué dirección es el impacto.

- "uplift_by_decile": Divide a la población en deciles (o cuantiles a determinar) basados en la predicción de un modelo de CATE en la muestra de entrenamiento, y devuelve el *uplift* promedio "real" para cada *outcome*, para cada cuantil. En la medida de lo posible, se trató de usar deciles, pero cuando el abanico de valores obtenidos por las predicciones continuas no lo permitían, se muestran los resultados para la mayor cantidad de cuantiles posibles.
- "uplift_by_value": Se utiliza cuando la predicción es una variable discreta con finitos valores. Divide a la población según los valores de la predicción, y devuelve el *uplift* promedio "real" para cada *outcome*, para cada valor.

Toda la comparación se hizo en Python. Las versiones utilizadas de las librerías de efectos causales son la versión 0.10.0 de CausalML y 0.9.0 de EconML.

Uplift para visit

Para empezar con la comparación, en las tablas siguientes se muestran los valores del error cuadrático medio ("Transformed Outcome Loss") en las muestras de entrenamiento y de testeo, para la campaña de mercadería femenina (tabla izquierda) mercadería masculina (tabla derecha). En la muestra de entrenamiento los errores son, en general más bajos, como es de esperarse, alcanzando el mínimo en ambas campañas en el caso del CRF (0.45 y 0.50) y tomando valores cercanos a 5 puntos porcentuales más altos para los otros modelos. En las muestra de testeo se observan valores más similares, cercanos a 0.52/0.53 en la campaña femenina y a 0.58 en la campaña masculina. Si bien en cada caso, no hay grandes diferencias entre los modelos, es interesante observar que en ambas campañas, el modelo superador en la muestra de entrenamiento es el que tiene mayor error en la muestra de testeo (CRF). Por otro lado, su versión honesta (HCRF) tenía mayor error que los RF alternativos en la muestra de entrenamiento, y sin embargo es el RF que presenta el menor error en la muestra de testeo. Esto parece sugerir una tendencia al *overfitting* por parte de los modelos CRF y DML HF.

Tabla 4-9: Comparación de MSE (TOL) para los modelos de 'visit'.

(a) Campaña Femenina

	TOL_train	TOL_test
ols	0.509103	0.526083
knn_275	0.508683	0.528393
hcrf	0.497737	0.526725
crf	0.451297	0.532966
dml_hf	0.487684	0.527729

(b) Campaña Masculina

	TOL_train	TOL_test
ols	0.569853	0.580196
knn_275	0.568903	0.580562
hcrf	0.557957	0.580366
crf	0.503882	0.587075
dml_hf	0.547075	0.581088

Claramente, la métrica anterior no es suficiente para comparar y elegir un modelo. Las figuras siguientes muestran las curvas Qini para las poblaciones de entrenamiento (panel izquierdo) y de testeo (panel derecho). Sólo se analizan las predicciones continuas; se excluye a la estimación obtenida por OLS en el caso de la campaña femenina ya que los finitos valores predichos impiden el correcto ordenamiento de la población según el score para calcular esta métrica, al igual que el AUUC. Al estudiar las curvas Qini, se prefieren los modelos cuyas curvas se alejan lo más posible de la curva de un modelo Random (diagonal negra).

En la muestra de entrenamiento, el orden de preferencia coincide con lo observado en la métrica MSE. Además, es evidente que el modelo de KNN y OLS (este último, solo analizado en la

campaña masculina) son inferiores que los modelos de RF, al mirar ambas muestras. En la muestra de testeo, en el caso de la campaña femenina, las curvas de los RF son bastante similares, siendo el HCRF (curva roja) el modelo superador, si acaso incluso el orden se invierte. En la campaña masculina, no se observa mayores diferencias entre las curvas; la única curva que se distingue y por debajo del resto, es la correspondiente a OLS.

Figura 4-20: Curvas Qini para los modelos de estimación continua para 'visit' – campaña femenina

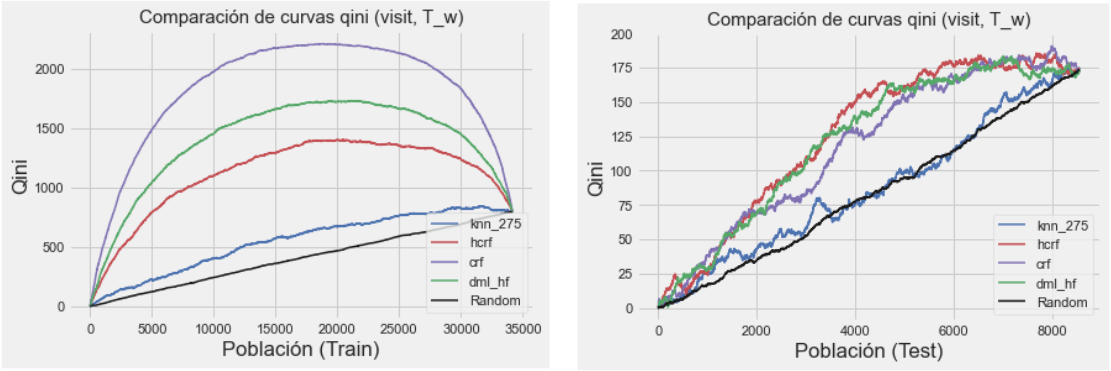
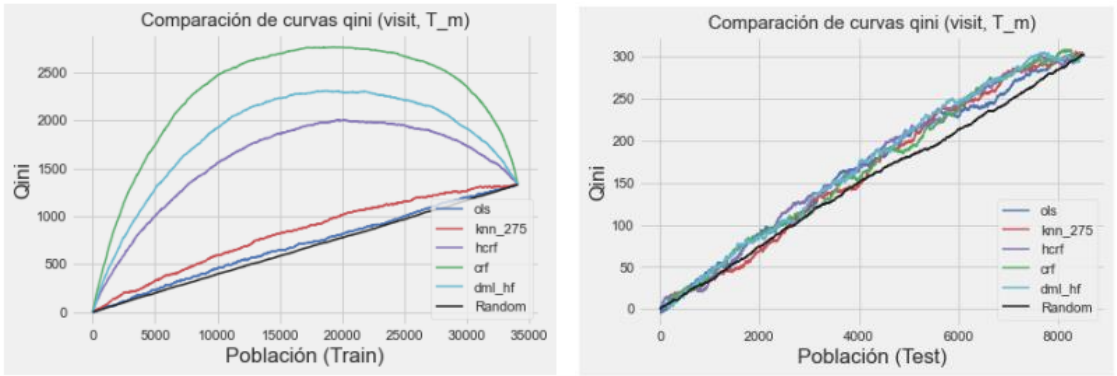


Figura 4-21: Curvas Qini para los modelos de estimación continua para 'visit' – campaña masculina



Los valores del Qini Score en la Tabla 4-10 resumen el comportamiento observado en los gráficos: el modelo de KNN y OLS (solo analizado en la campaña masculina) son inferiores que los modelos de RF, y en la muestra de testeo los valores son menores, especialmente para la campaña masculina.

Se comentó anteriormente que la muestra de datos de Hillstrom es popular a la hora de aplicar y probar metodologías de predicción de efectos causales para datos reales, aunque no se estudian todas las variantes de campaña y *outcome*, y se concentran en el *uplift* para 'visit' en la campaña femenina. En ese caso en particular, cabe recalcar que los valores obtenidos en la muestra de entrenamiento para los RF están en el rango de los resultados en el trabajo de Devriendt et al. [7] (ver tabla 6 de dicho trabajo), entre 39,53% y 50,50%. Lamentablemente, para la campaña masculina como para los otros *outcomes*, no se cuenta con valores de referencia.

Tabla 4-10: Comparación del Qini para los modelos de 'visit'.

(a) Campaña Femenina

	qini_train	qini_test
knn_275	17.57%	2.45%
hcrf	43.18%	23.98%
crf	46.98%	19.43%
dml_hf	45.26%	22.17%

(b) Campaña Masculina

	qini_train	qini_test
ols	2.79%	4.67%
knn_275	11.97%	3.74%
hcrf	40.80%	6.10%
crf	46.09%	4.95%
dml_hf	43.48%	6.33%

En el caso de la métrica AUUC, se observa un patrón similar. El modelo KNN no es mucho mejor que un modelo random en la muestra de entrenamiento en ambas campañas, y en el caso de la campaña masculina, el AUUC para OLS en la muestra de entrenamiento es apenas 52%. En la muestra de testeo, en el caso de la campaña de mercadería masculina todos los modelos tienen un AUUC bajo, cercano a 50%, mientras que en la campaña femenina, los RF presentan valores más altos. En general, todos los modelos muestran evidencia de *overfitting*, especialmente el CRF, que tiene el mayor AUUC en la muestra de entrenamiento y el menor (después de KNN) en la muestra de testeo. El modelo HCRF es el RF que achica el gap de la métrica en ambas muestras: tiene el menor AUUC en la muestra de entrenamiento pero el mayor (por decimales, el segundo mayor en la campaña masculina) en la muestra de testeo.

Tabla 4-11: Comparación de AUUC ("Area under the uplift Curve") para los modelos de estimación continua para 'visit'.

(a) Campaña Femenina

	auuc_train	auuc_test
knn_275	68.19%	50.14%
hcrf	93.44%	71.35%
crf	97.08%	66.72%
dml_hf	95.45%	69.52%

(b) Campaña Masculina

	auuc_train	auuc_test
ols	52.73%	55.38%
knn_275	61.95%	54.66%
hcrf	90.54%	57.02%
crf	95.29%	55.45%
dml_hf	93.00%	57.09%

Cabe recordar que uno de los beneficios de considerar la cualidad de honestidad era el de reducir el *overfitting*. Esto se verifica en este caso, al comparar HCRF y CRF, que son modelos con las mismas especificaciones salvo por la honestidad. En líneas generales, en base a las métricas de performance consideradas, los modelos de OLS y KNN son los que muestran mayores deficiencias, siendo superado en cada caso por los modelos de RF, entre los cuales el HCRF parece ser el superador en base a *overfitting* y comportamiento de las métricas en la muestra de testeo.

Con el objetivo de elegir un modelo, además de mirar conjuntamente las métricas anteriores, también es útil comparar las estimaciones en su relación con las variables predictoras y habilidad de *rankear* correctamente a la población, como se analiza a continuación.

Para lo primero, se utilizó la función "pred_CATE_by_bin" definida ad hoc (referirse al comienzo de la sección 4.4) y se muestran los resultados para las variables identificadas como más importantes en el *uplift* para 'visit' (ver Tabla 4-3 y Tabla 4-6).

Para la campaña femenina, los resultados se encuentran en el Apéndice III:

- Salvo KNN, todos los modelos son consistentes al indicar que el *uplift* es mayor en los individuos con "womens =1" (Tabla 8-1) y "mens =0" (Tabla 8-3). Es decir: la población que sólo compro mercadería femenina en el pasado reacciona mejor a esta campaña de mercadería femenina. Cabe recordar que 'mens' es la única variable utilizada en el modelo de OLS, y entonces es esperable que distinga el impacto para esta variable. Sin embargo, también distingue bien en el caso de 'womens' ya que, aunque estas variables no son complementarias, sólo hay un 10% de observaciones que compraron de ambas mercaderías en el pasado[11].

[11] Concretamente, en total en ambas campañas hay 6448 clientes que compraron tanto mercadería femenina como masculina, 28828 compraron sólo mercadería masculina y 28734 compraron sólo mercadería femenina.

- Salvo OLS, todos los modelos coinciden en que el mayor *uplift* se observa para los segmentos de gasto 5, 6 y 7, y que el impacto en los segmentos 3 y 4 es menor que en los segmentos 1 y 2 (Tabla 8-2).

- A grandes rasgos, los modelos de RF concluyen consistentemente que para los valores de 'recency' de 7, 8, 9 y 12 meses se observa mayor *uplift,* y la diferencia es marginalmente más notoria en el caso de CRF. En cambio, no se observan mayores variaciones en las predicciones de KNN y OLS. (Tabla 8-4)

- En el análisis de importancia de las variables la dummy 'zip_code_Rural' aparecía en quinto lugar. Lo que se observa en la Tabla 8-5, y de manera más pronunciada en HCRF, es que el impacto de la campaña es mayor para 'zip_code' "Surburban" y "Urban", que es justamente cuando 'zip_code_Rural=0'. En cambio, no se observan mayores variaciones en las predicciones de KNN y OLS.

Para la campaña masculina, los resultados se encuentran en el Apéndice IV:

- Como el modelo de OLS usa la variable 'history' y con un coeficiente de signo positivo, el *uplift* incrementa a medida que aumenta el gasto histórico, y este modelo no capta comportamientos distintos para distintos segmentos de gasto. Salvo OLS, todos los modelos coinciden en que el mayor *uplift* se observa para los segmentos de gasto 3, 4, 5 y 7. (Tabla 9-1).

- A diferencia de lo observado en la campaña femenina, en este caso, en general se observa mayor *uplift* para menores valores de 'recency'. Las variaciones son más visibles en los RF y, en menor medida, en KNN. Dado que 'recency' no ingresó en OLS, este modelo prácticamente no distingue los distintos valores. (Tabla 9-2).

- Los RF son consistentes al indicar que el *uplift* es marginalmente mayor en los individuos con "womens =1" (Tabla 9-3), mientras que el efecto para ambos valores de 'mens' es similar. Es decir: la población que compró mercadería femenina en el pasado, independientemente de si haya comprado o no mercadería masculina, reacciona mejor a esta campaña de mercadería masculina.

- En el análisis de importancia de las variables la dummy 'zip_code_Rural' aparecía en quinto lugar. Si bien el efecto para los valores de 'zip_code' son similares, al igual que en la campaña femenina, se observa que el impacto de la campaña es marginalmente mayor para 'zip_code' "Surburban" y "Urban", que es justamente cuando "zip_code_Rural=0". (Tabla 9-4)

En resumen, se puede observar que los modelos detectan de manera consistente los valores o rangos de las variables que mayor impacto tienen en el *uplift* de 'visit', aunque en general las variaciones son más notorias en los RF. OLS sólo distingue los valores de las variables predictoras si fueron incluidas en el modelado ('womens' en el caso de la campaña femenina, y 'history' en el caso de la campaña masculina).

Teniendo en cuenta que el objetivo de modelar el efecto de tratamiento es optimizar la campaña, a continuación, todavía resta utilizar los modelos para identificar a los individuos más propensos a reaccionar de manera positiva. Para esto, se utilizó la función "uplift_by_decile" definida ad hoc (referirse al comienzo de la sección 4.4). Dado que optimizar la campaña tiene tres interpretaciones, para cada modelo de *uplift* en las visitas al sitio, se analiza cómo los modelos ordenan a los individuos tanto en el *outcome* modelado ('visit'), sino también en 'conversion' y 'spend'. En este análisis, nuevamente, se excluye a la estimación obtenida por OLS en el caso de la campaña femenina, ya que los finitos valores predichos impiden el correcto ordenamiento de la población según el *score*. Para dicho caso en particular, sólo se puede comparar el *uplift* promedio predicho para ambos valores de predicciones posibles. El modelo distingue dos grupos, y para el de mayor *uplift* en visit (columna derecha), la campaña tuvo mayor impacto en los tres *outcomes* de interés. Si bien esto es lo deseado, una distinción tan limitada dentro de la población,

limita el análisis y las decisiones posteriores de elegir una población a la cual dirigir una futura campaña[12].

Tabla 4-12: Ranking de la población total según valores de visit_ols – campaña Femenina

	0.025691	0.071868
visit	2.18%	7.40%
conversion	0.15%	0.51%
spend	0.256416	0.631648
#	23527	19166

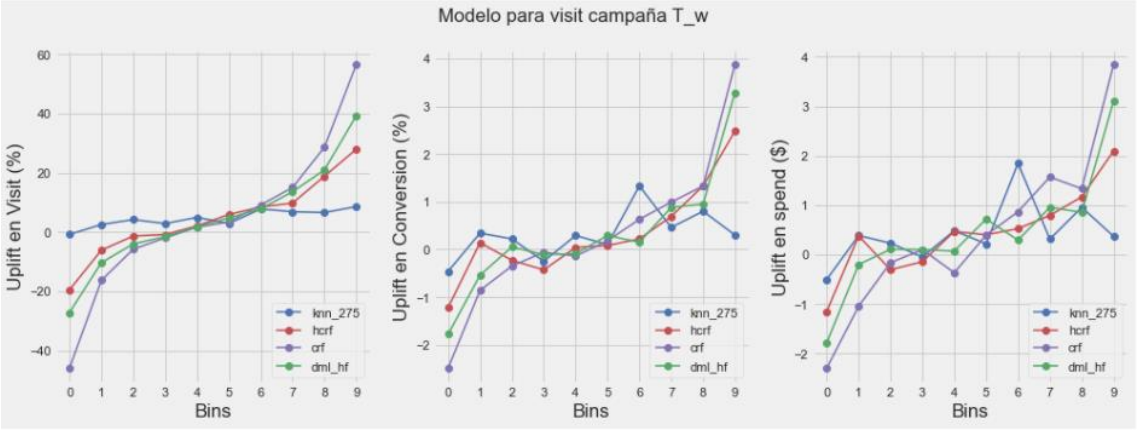
En las figuras siguientes se ve que, salvo OLS para la campaña masculina, los modelos graficados ordenan correctamente a la población según el impacto en las visitas al sitio (panel de la izquierda) para ambas campañas. Sin embargo, la discriminación definida por el modelo de KNN no es la deseada: la curva es prácticamente horizontal e incluso presenta algunas caídas leves. En el caso de OLS, además pareciera que el primer cuantil (el de menor *score*) presenta mayor *uplift*; esto es lo opuesto a lo deseado. En cambio, los modelos de RF presentan curvas monótonas crecientes con pendientes mayores, y cubriendo un mayor rango de valores en el eje y, incluso valores negativos. Para mayor detalles sobre el rango de valores de las predicciones, los gráficos en la Figura 8-1 en el Apéndice III y Figura 9-1 en el Apéndice IV muestran la distribución de las predicciones continuas consideradas para el *uplift* de la variable 'visit'.

Uplift negativo esencialmente describe un efecto adverso de la campaña en el público. Un modelo ideal es aquel que ordena a la población de tal manera que los individuos de *score* más alto son más propensos a reaccionar positivamente a la campaña. Por el contrario, los *scores* negativos pretenden identificar a aquellos individuos que no visitaron el sitio como consecuencia de la campaña, en este caso por ejemplo, en comparación con individuos de similares características pero en el grupo de control que sí hubieran visitado el sitio.

Las curvas de *uplift* en la cantidad de compras (panel central) y el monto en dólares de dichas compras (panel de la derecha) son un poco más erráticas, especialmente en la campaña femenina, incluso en los modelos de RF. Esto es lógico, ya que no son el *outcome* modelado en este caso. De todas maneras, desde el punto de vista del negocio, no hay que dejar de mirar dicho comportamiento, ya que en última instancia, el objetivo es aumentar las comprar y las ganancias. Mientras que los modelos de KNN siguen mostrando deficiencias, los modelos de RF ordenan de manera aceptable, por lo menos en los últimos cuantiles, salvo CRF en el caso de 'spend' de la campaña femenina, que presenta un *uplift* mayor en el decil 7[13] que en el 8. En cualquier caso, si se tuviera que elegir un 10% de la población, eligiendo el último cuantil según cualquier RF logra identificar a individuos que mejor reaccionan a las campañas en base a los tres *outcomes* de interés. Esto se comentará en ms detalle en las conclusiones finales de la sección 5, comparando con el valor de referencia del ATE para cada campaña completa.

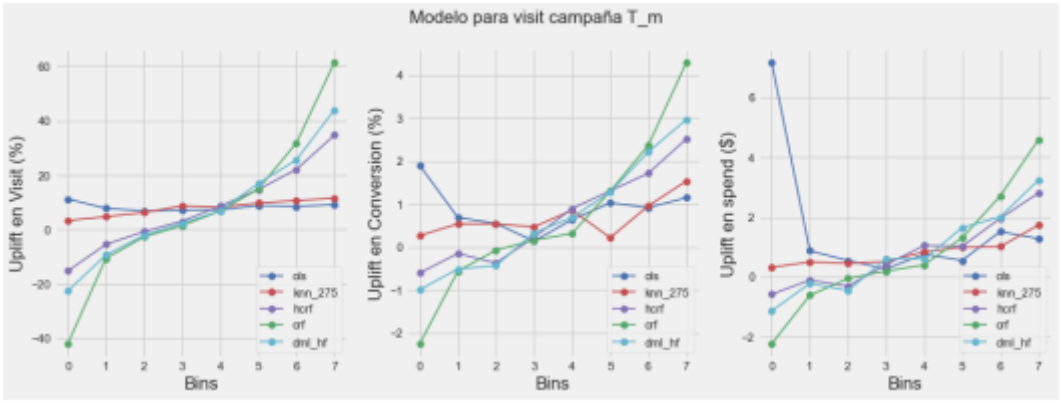
[12] Es cierto que se hubiera podido elegir otra variable predictora con más valores, como en el caso de la campaña masculina. Sin embargo, como se verá en ese caso, OLS sigue siendo un modelo deficiente.
13 Notar que la numeración de los deciles comienza en 0 y termina en 9.

Figura 4-22: Ranking de la población total según deciles – campaña Femenina



Nota: Los deciles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (WOMEN_campaign)

Figura 4-23: Ranking de la población total según cuantiles – campaña Masculina



Nota: Los cuantiles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (MEN_campaign)

La tabla siguiente muestra el *uplift* en cada *outcome* para los modelos de 'visit', para el último decil de la población total de la campaña femenina, según cada modelo. Además se muestra como valor de referencia el ATE. Mientras todos los modelos logran superar el efecto promedio en las visitas al sitio al seleccionar solo un 10% de la población total, sólo los RF logran superar el *uplift* en los otros *outcomes* de interés. OLS no es comparable, ya que esencialmente divide a la población en dos. Este modelo logra identificar a una mitad de la población de tal manera de lograr un *uplift* en las ventas apenas superior al promedio total, sí superando al valor de KNN seleccionando 10% de la población, pero ampliamente superado por los RF.

Tabla 4-13: *Uplift* en cada *outcome* de interés, para el último decil de WOMEN_CAMPAIGN

WOMEN (deciles)	visit (%)	conversion (%)	spend ($ per capita)
ATE	4.52	0.31	0.42
KNN	8.59	0.3	0.37
CRF	56.66	3.88	3.85
HCRF	27.98	2.48	2.1
DML HF	39.33	3.29	3.11

OLS*	7.4	0.51	0.63

Nota: * Los valores de OLS corresponden al grupo con $\widehat{CATE}(X) = 0.071868$ (Tabla 4-12)

En el caso de la campaña masculina, siendo el más deficiente OLS, se lo excluye del análisis a continuación, para poder dividir a la población en deciles al igual que en la campaña femenina.

Tabla 4-14: *Uplift* en cada *outcome* de interés, para el último decil de MEN_CAMPAIGN

MEN (deciles)	visit (%)	conversion (%)	spend ($ per capita)
ATE	7.66	0.68	0.77
KNN	13.16	1.78	2.19
CRF	65.17	4.72	5.08
HCRF	36.27	2.58	3.09
DML HF	46.17	3.25	3.48

Si bien los tres *outcomes* de interés indican consistentemente que la campaña de mercadería masculina es la más exitosa, en proporción, los modelos de la campaña femenina parecen identificar mejor a los mejores individuos. Es decir, en la campaña masculina es más marcada la necesidad de dirigir la campaña a una mayor proporción de clientes para aumentar el *uplift*. Devriendt et al. [7] explica que esto suele indicar que en la población no hay gran proporción de los llamados "do-not-disturbs". Este tipo de clientes son aquellos que hubieran respondido si no hubieran sido tratados, pero no responden si fueron tratados. Explica también que esto se manifiesta en curvas Qinis más indiferenciadas de la Random, comportamiento llamativo en la muestra de testeo de la campaña masculina.

A continuación se muestra un análisis similar, pero al modelar directamente el *uplift* en 'conversion' y en 'spend'.

Uplift para conversion

En cuanto a MSE, las tablas a continuación muestran que, al igual que en los modelos de 'visit', el modelo superador en la muestra de entrenamiento tiene mayor error en la muestra de testeo (CRF), sólo superado por OLS en el caso de la campaña femenina. Se puede observar que los modelos se comportan de manera similar, con errores pequeños. Incluso no hay grandes diferencia entre las muestra de entrenamiento y de testeo. Errores pequeños son, de hecho, esperables en este caso y no es necesariamente evidencia de un modelo preciso. Los datos son muy desbalanceados en términos de los valores del *outcome* de estudio ('conversion'). Como se comentó en la sección 3.5 (Tabla 3-6 y Tabla 3-7) la proporción de compras en la campaña femenina es del 0.73% mientras que para la campaña masculina es de 0.91% y el ATE 0.31% y 0.68%, respectivamente (Tabla 3-5). A grandes rasgos, esto quiere decir que un modelo simple que asigna a cada observación una predicción de *uplift* nula sería bastante bueno.

Tabla 4-15: Comparación de MSE (TOL) para los modelos de 'conversion'.

(a) Campaña Femenina (b) Campaña Masculina

	TOL_train	TOL_test
ols	0.028940	0.031096
knn_150	0.028573	0.030905
hcrf	0.027973	0.030920
crf	0.026267	0.031026
dml_hf	0.027284	0.030973

	TOL_train	TOL_test
ols	0.036908	0.034688
knn_200	0.036809	0.034760
hcrf	0.035964	0.034774
crf	0.033610	0.035163
dml_hf	0.035164	0.034822

Las figuras siguientes muestran las curvas Qini para las poblaciones de entrenamiento (panel izquierdo) y de testeo (panel derecho) de los modelos con predicciones continuas. Nuevamente, es evidente que el modelo de KNN, y OLS en la campaña masculina, son notoriamente inferiores a los modelos de RF en la muestra de entrenamiento. En la muestra de testeo, las curvas son bastante erráticas; en la campaña femenina la que más se aleja de la diagonal es la curva correspondiente a KNN, seguida por HCRF. En la campaña masculina, todos los modelos parecen deficientes, especialmente KNN y CRF.

Figura 4-24: Curvas Qini para los modelos de estimación continua para 'conversion' – campaña Femenina

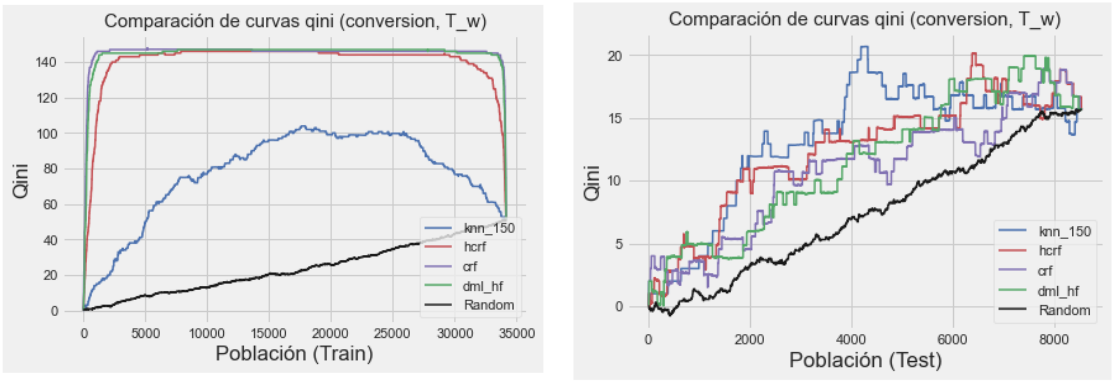
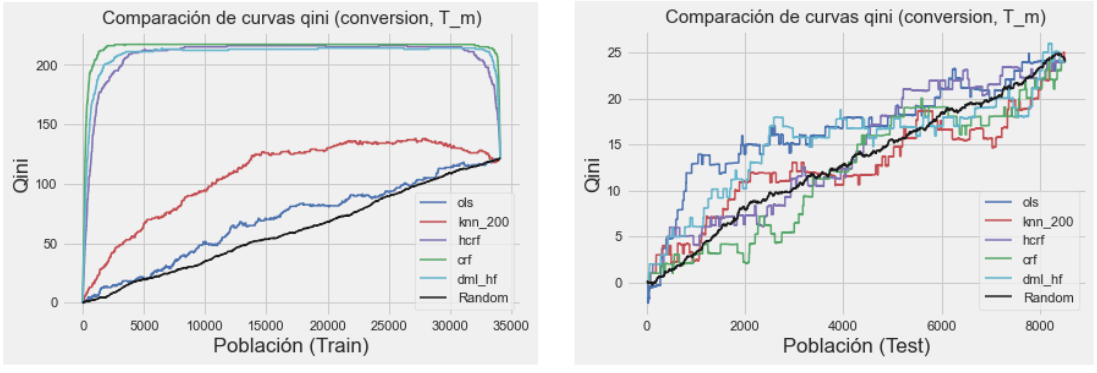


Figura 4-25: Curvas Qini para los modelos de estimación continua para 'conversion' – campaña Masculina



Nuevamente, los valores del Qini Score resumen el comportamiento observado en los gráficos: el modelo de KNN y OLS (solo analizado en la campaña masculina) son inferiores que los modelos de RF en la muestra de entrenamiento, y en la muestra de testeo los valores son menores. En la campaña masculina se observan incluso valores negativos, dado que las curvas de KNN y CRF encierran mucha área por debajo de la curva Random.

Tabla 4-16: Comparación del Qini Score para los modelos de 'conversion'.

(a) Campaña Femenina

	qini_train	qini_test
knn_150	46.14%	32.64%
hcrf	52.87%	29.03%
crf	53.25%	20.23%
dml_hf	53.17%	22.07%

(b) Campaña Masculina

	qini_train	qini_test
ols	7.32%	15.14%
knn_200	33.79%	-1.35%
hcrf	49.24%	5.10%
crf	49.78%	-4.01%
dml_hf	49.59%	8.35%

Como se comentó en la sección 2.1.4, cuando se trata de datos con *outcome* desbalanceado es preferible analizar la performance de un modelo con el AUUC [14]. En la muestra de entrenamiento, los modelos de RF son superiores a KNN y a OLS en la campaña masculina, aunque todos los modelos muestran valores llamativamente favorables, salvo OLS. En general, todos los modelos muestran evidencia de *overfitting* al comparar con los valores en la muestra de testeo, especialmente el CRF, que tiene el mayor AUUC en la muestra de entrenamiento y el menor en la muestra de testeo.

Tabla 4-17: Comparación de AUUC ("Area under the uplift Curve") para los modelos de estimación continua para 'conversion'.

(a) Campaña Femenina

	auuc_train	auuc_test
knn_150	92.82%	79.38%
hcrf	99.38%	76.47%
crf	99.78%	66.88%
dml_hf	99.71%	69.35%

(b) Campaña Masculina

	auuc_train	auuc_test
ols	57.31%	69.37%
knn_200	83.50%	52.82%
hcrf	99.02%	59.91%
crf	99.54%	50.69%
dml_hf	99.30%	62.77%

Al igual que lo hecho para 'visit', a continuación se comparan las estimaciones en términos de su relación con las variables predictoras identificadas como más importantes en el *uplift* para 'conversion' (ver Tabla 4-4 y Tabla 4-7).

Para la campaña femenina, los resultados se encuentran en el Apéndice III:

- Todos los modelos coinciden en que el mayor *uplift* se observa para los segmentos de gasto 5, 6 y 7, y que el impacto en los segmentos 3 y 4 es menor (incluso, negativo) que en los segmentos 1 y 2 (Tabla 8-6). Vale la pena recordar que lo mismo se podía observar en los modelos para 'visit'.
- En general, no se observan mayores variaciones pero a grandes rasgos, los modelos de RF concluyen consistentemente que para los valores de 'recency' 3 y 8 meses se observa menor *uplift,* y la diferencia es marginalmente más notoria en el caso de CRF. En cambio, las predicciones de KNN se encuentran cercanas a 0.30% para todos los valores de 'recency' (Tabla 8-7).
- Al igual que lo concluido a partir de los modelos de 'visit', los RF son consistentes al indicar que la población que sólo compró mercadería femenina en el pasado reacciona mejor a esta campaña de mercadería femenina (Tabla 8-8 y Tabla 8-9).

Para la campaña masculina, los resultados se encuentran en el Apéndice IV:

- Todos los modelos coinciden en que en general, el *uplift* es mayor para segmentos de mayor gasto, excepto el segmento 5 (Tabla 9-6). Como el modelo de OLS usa la variable 'history' y con un coeficiente de signo positivo, el *uplift* incrementa a medida que aumenta el gasto histórico, y este modelo no capta un comportamiento distinto para el segmento 5.
- Se observa que el impacto de la campaña es mayor para 'zip_code' "Surburban" y "Urban", que es justamente cuando "zip_code_Rural=0". Las variaciones son mínimas para KNN y nulas para OLS. (Tabla 9-7)

A la hora de analizar el correcto ordenamiento de los modelos para 'conversion' según los *outcomes* de interés, esta vez se dividió a la población en quintiles, ya que el abanico de valores de algunos modelos (knn_150 y CRF para la campaña femenina, y OLS y knn_200 para la campaña masculina) no permitía distinguir 10 intervalos distintos para dichas predicciones.

Además, para el caso de OLS en la campaña femenina, en la tabla a continuación sólo se puede comparar el *uplift* promedio predicho para ambos valores de predicciones posibles. El modelo divide a la población en los mismos dos grupos que el modelo para *uplift* en visit.

Tabla 4-18: Ranking de la población total según valores de conversion_ols – campaña Femenina

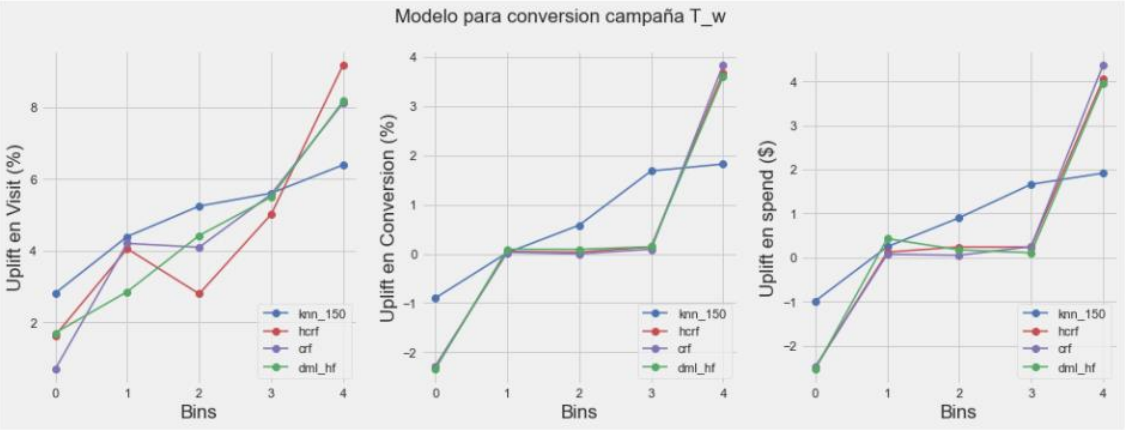
	-0.019673	0.004297
visit	2.18%	7.40%
conversion	0.15%	0.51%
spend	0.256416	0.631648
#	23527	19166

En las figuras siguientes se ve que los tres RF no ordenan o no distinguen satisfactoriamente los quintiles centrales. Sin embargo, de alguna manera los modelos distinguen tres grupos bien ordenados, y el último quintil logra identificar a individuos que mejor reaccionan a la campaña en base a los tres *outcomes* de interés. El KNN, por otro lado, presenta curvas monótonas crecientes para los 'conversion' y 'spend'[14]. El modelo de OLS para la campaña masculina sigue mostrando deficiencias con curvas principalmente horizontales. Más aún, ambos extremos de las curvas de 'conversion' y 'spend' están a la misma altura; es decir, no distingue los individuos que peor reaccionaron a la campaña de los que mejor reaccionaron. Al igual que lo observado en 'visit', los modelos de RF cubren un mayor rango de valores en el eje y.

Para mayor detalles sobre el rango de valores de las predicciones, los gráficos en la Figura 8-2 en el Apéndice III y Figura 9-2 en el Apéndice IV muestran la distribución de las cuatro predicciones continuas consideradas para el *uplift* de la variable 'conversion'.

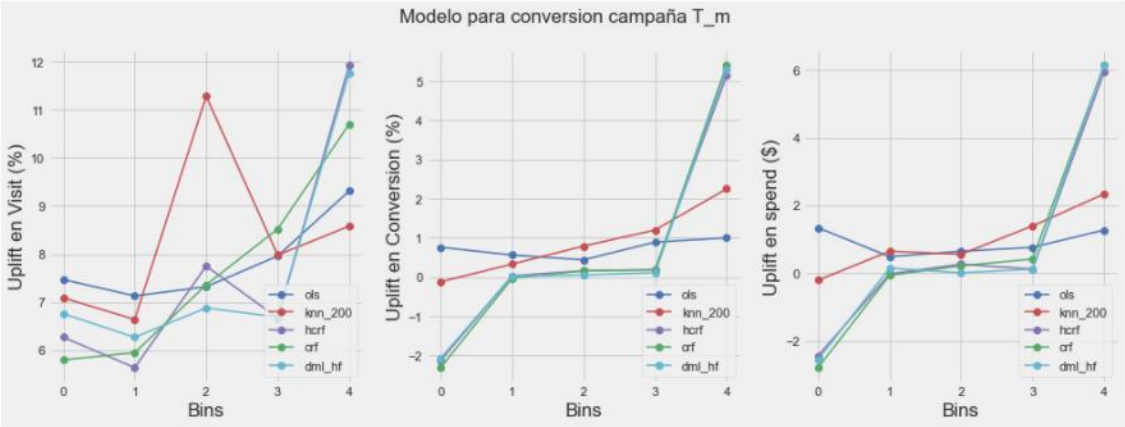
[14] En el caso de la campaña femenina, también ranquea correctamente en 'visit'. Sin embargo, dado que lo que se modeló en esta oportunidad es el impacto en la cantidad de compras, se prioriza el análisis en las instancias posteriores a la visita al sitio (compra y gasto).

Figura 4-26: Ranking de la población total según quintiles – Campaña Femenina



Nota: Los quintiles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (WOMEN_campaign)

Figura 4-27: Ranking de la población total según quintiles – Campaña Masculina



Nota: Los quintiles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (MEN_campaign)

Uplift en spend

En cuanto a MSE, la tabla a continuación muestra que todos los modelos presentan errores altos, y más aún en la muestra de testeo de la campaña femenina. Se arrastra el problema identificado en los modelos de 'conversion', donde se contaba con una población muy desbalanceada; ahora la distribución de la variable 'spend' es muy asimétrica (ver Figura 3-2 y Figura 3-3). En el caso de los modelos de 'conversion' los errores eran pequeños ya que habiendo sólo dos valores posibles para dicho *outcome*, se podía pensar en un modelo simple de bastante precisión, lo cual no es posible en este caso, dada la dispersión de la variable 'spend'.

Tabla 4-19: Comparación de MSE (TOL) para los modelos de 'spend'.

(a) Campaña Femenina (b) Campaña Masculina

	TOL_train	TOL_test
ols	634.053425	1108.076764
knn_150	631.746988	1107.054224
hcrf	617.659111	1108.635139
crf	586.263390	1112.305567
dml_hf	603.310994	1108.773773

	TOL_train	TOL_test
ols	906.914865	888.265306
knn_200	905.369010	889.112832
hcrf	883.622105	890.923354
crf	834.598615	896.558453
dml_hf	863.459374	893.667680

Las figuras siguientes muestran las curvas Qini para las poblaciones de entrenamiento (panel izquierdo) y de testeo (panel derecho) de los modelos con predicciones continuas, para ambas campañas. Se observa el mismo comportamiento que en los modelos de 'visit' y 'conversion': el modelo de KNN es notoriamente inferior a los modelos de RF en la muestra de entrenamiento. En la muestra de testeo, las curvas son bastante erráticas siendo en la campaña femenina la que más se aleja de la diagonal la curva correspondiente a KNN, seguida por HCRF. En el caso de la campaña masculina, CRF y DML_HF parecen ser superadores. OLS es el más deficiente en la muestra de entrenamiento.

Figura 4-28: Curvas Qini para los modelos de estimación continua para 'spend'- campaña Femenina

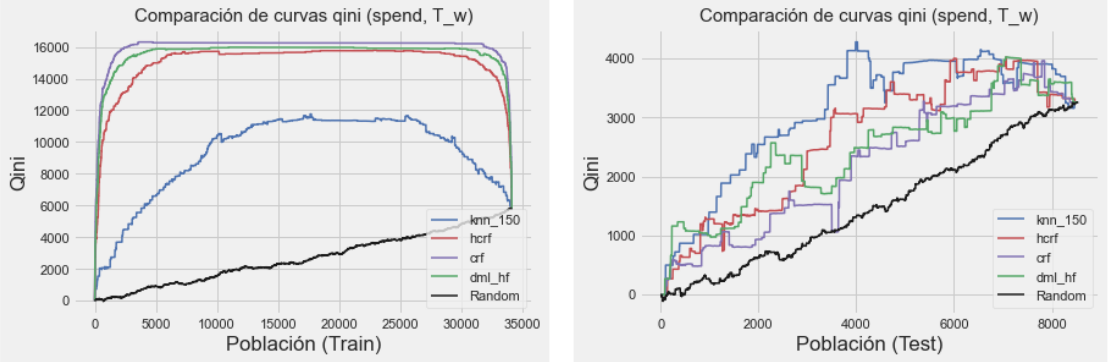
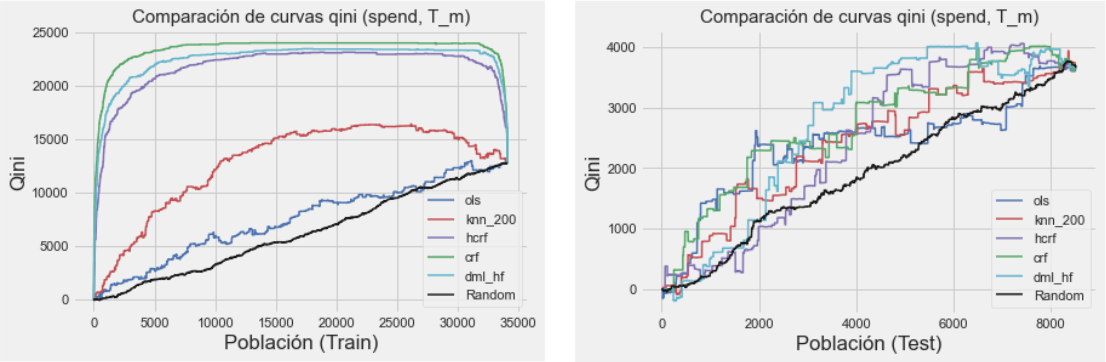


Figura 4-29: Curvas Qini para los modelos de estimación continua para 'spend' – campaña Masculina



Nuevamente, los valores del Qini Score de las tablas a continuación resumen el comportamiento observado en los gráficos.

Tabla 4-20: Comparación del Qini Score para los modelos de 'spend'.

(a) Campaña Femenina (b) Campaña Masculina

	qini_train	qini_test
knn_150	49.12%	38.52%
hcrf	53.32%	29.63%
crf	53.50%	19.32%
dml_hf	53.44%	26.29%

	qini_train	qini_test
ols	9.87%	12.58%
knn_200	38.11%	12.86%
hcrf	50.74%	12.13%
crf	51.24%	20.47%
dml_hf	51.08%	19.61%

Al igual que para 'conversion', y al tratarse de datos con *outcome* desbalanceado es preferible analizar la performance de un modelo con el AUUC [14]. Nuevamente, en la muestra de entrenamiento, los modelos de RF son superiores a KNN, y a OLS en el caso de la campaña masculina, aunque todos los modelos muestran valores llamativamente favorables. En general, todos los modelos muestran evidencia de *overfitting* al comparar con los valores en la muestra de testeo.

Tabla 4-21: Comparación de AUUC ("Area under the uplift Curve") para los modelos de estimación continua para 'spend'.

(a) Campaña Femenina

	auuc_train	auuc_test
knn_150	95.67%	83.82%
hcrf	99.77%	75.75%
crf	99.87%	65.04%
dml_hf	99.86%	72.43%

(b) Campaña Masculina

	auuc_train	auuc_test
ols	58.51%	64.42%
knn_200	86.60%	64.99%
hcrf	99.30%	64.21%
crf	99.69%	73.15%
dml_hf	99.60%	71.93%

Las variables predictoras identificadas como más importantes en el *uplift* para 'spend' (ver Tabla 4-5 y Tabla 4-8) son 'history' y 'recency' en ambas campañas.

Para la campaña femenina, los resultados se encuentran en el Apéndice III:

- Todos los modelos coinciden en que el mayor *uplift* se observa para los segmentos de gasto 5, 6 y 7, y que el impacto en los segmentos 3 y 4 es menor (incluso, negativo) que en los segmentos 1 y 2. Vale la pena recordar que lo mismo se podía observar tanto en los modelos para 'visit' como para 'conversion' (Tabla 8-10).
- En cuanto a la antigüedad de la última compra, los modelos concluyen que para los valores de 'recency' 3, 7 y 8 meses se observa menor *uplift*, mientras que para los meses 4 y 5 el impacto de la campaña es mayor (Tabla 8-11).
- Todos los modelos concluyen consistentemente que la campaña es más efectiva dentro de los clientes nuevos (Tabla 8-12).

Para la campaña masculina, los resultados se encuentran en el Apéndice IV:

- Como el modelo de OLS usa la variable 'history' y con un coeficiente de signo positivo, el *uplift* incrementa a medida que aumenta el gasto histórico, y este modelo no capta comportamientos distintos para distintos segmentos de gasto. Salvo OLS, todos los modelos coinciden en que el mayor *uplift* se observa para los segmentos de gasto 4, 6 y 7, similar a lo observado para 'visit' (Tabla 9-8).

- En cuanto a la antigüedad de la última compra, mientras que el KNN no muestra grandes variaciones, los modelos concluyen que para los valores de 'recency' 3 y 12 meses se observa mayor *uplift,* mientras que para el mes 8 el impacto de la campaña es menor (Tabla 9-9).
- Todos los modelos concluyen consistentemente que la campaña es más efectiva dentro de los clientes nuevos, aunque de manera poco acentuada para KNN (Tabla 9-10).
- Se observa que el impacto de la campaña es mayor para 'Channel' igual a "Multichannel" y luego "Web", que es justamente cuando "Channel_Phone=0". Es decir, la campaña es menos efectiva en los clientes que en el pasado sólo compraron via teléfono. (Tabla 9-11)

En cuanto al correcto ordenamiento de la población, nuevamente para el caso de OLS en la campaña femenina, en la tabla a continuación sólo se puede comparar el *uplift* promedio predicho para ambos valores de predicciones posibles. El modelo divide a la población en los mismos dos grupos que el modelo para *uplift* en 'visit' y 'conversion'.

Tabla 4-22: Ranking de la población total según valores de spend_ols – campaña Femenina

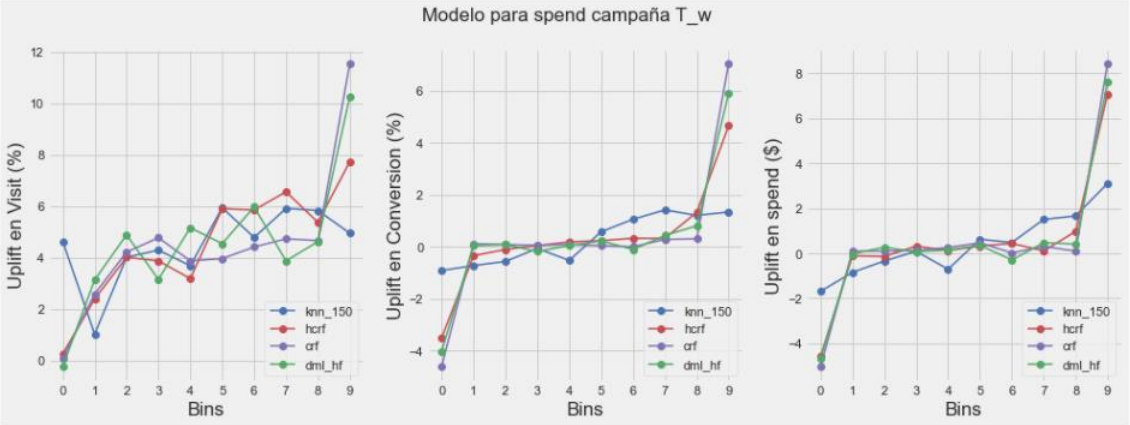
	0.255434	0.444164
visit	2.18%	7.40%
conversion	0.15%	0.51%
spend	0.256416	0.631648
#	23527	19166

En las figuras siguientes se ve que los modelos de *uplift* para 'spend' no ordenan satisfactoriamente según el impacto en términos de visitas al sitio (panel de la izquierda), como ya ocurrió en los modelos de 'conversion'. En cuanto a los otros *outcomes* de interés, similar a lo observado en los modelos de 'conversion', los tres RF no ordenan o distinguen satisfactoriamente los cuantiles centrales pero, de alguna manera los modelos distinguen tres grupos bien ordenados, y el último cuantil logra identificar a individuos que mejor reaccionan a la campaña en base a las compras y sus montos. La curva de KNN, por otro lado, si bien presenta mayor pendiente en los cuantiles centrales, con algunas caídas, cubre un rango de valores en el eje y más limitado que los RF, los cuales distinguen mejor los individuos con *uplift* negativo.

El modelo de OLS para la campaña masculina sigue mostrando deficiencias con curvas principalmente horizontales con extremo inicial más alto que el extremo final; es decir, no distingue los individuos que peor reaccionaron a la campaña de los que mejor reaccionaron.

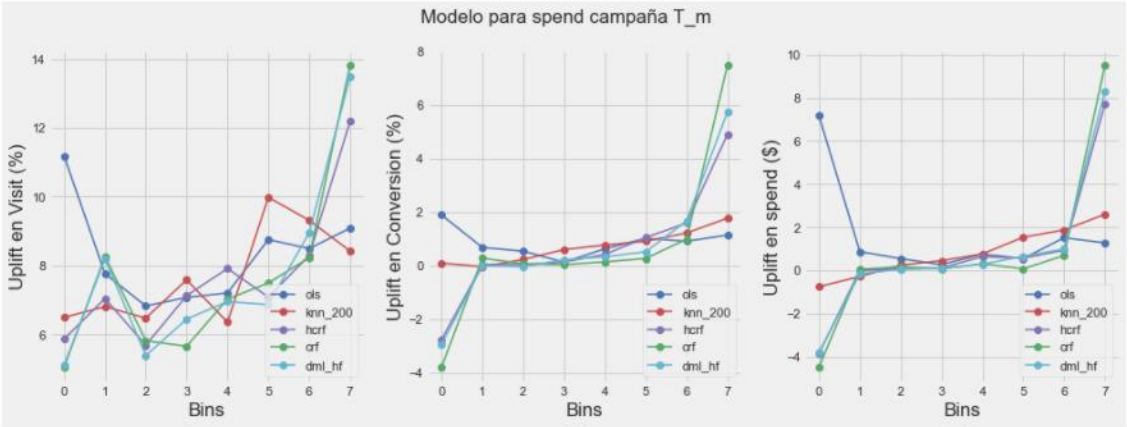
Para mayor detalles sobre el rango de valores de las predicciones, los gráficos en la Figura 8-3 en el Apéndice III y Figura 9-3 en el Apéndice IV muestran la distribución de las cuatro predicciones continuas consideradas para el *uplift* de la variable 'spend'.

Figura 4-30: Ranking de la población total según deciles – campaña Femenina



Nota: Los deciles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (WOMEN_campaign)

Figura 4-31: Ranking de la población total según cuantiles – campaña Masculina



Nota: Los cuantiles están calculados a partir de la población de entrenamiento (para cada modelo) y aplicado a la población total (MEN_campaign)

5 Conclusiones

En línea con los objetivos generales propuestos en esta tesis, se compararon varios modelos para el análisis de efectos de tratamiento en campañas de marketing, entre ellos: regresiones lineales, vecinos más cercanos y tres tipos de Random Forests.

A la hora de evaluar y comparar los modelos, es importante tener presente dos objetivos específicos de este trabajo: estimar el efecto de tratamiento y ordenar a la población. Siendo que el verdadero efecto de tratamiento es desconocido, el primer objetivo pasa a un segundo plano mientras que utilizar un modelo para el ordenamiento de la población permite identificar a los individuos más receptivos para futuras campañas. Es por eso que se evaluaron conjuntamente las métricas y la capacidad de ordenamiento.

En cuanto a las métricas, los modelos para *uplift* de 'conversion' y 'spend' muestran ciertas deficiencias. El MSE no provee demasiada información para evaluar la *performance*, tomando valores muy pequeños y parecidos dentro de cada campaña para los modelos de 'conversion', y muy grandes para 'spend'. En términos del Qini, los RF son superiores en la muestra de entrenamiento, pero al analizar la muestra de testeo, las curvas presentan un comportamiento errático, y un modelo superador no es obvio. Al comparar los modelos de *uplift* para 'visit', las métricas en general son un poco más concluyentes a favor de los RF. CRF es el RF superador en las muestras de entrenamiento pero al observar las muestras de testeo, es superado tanto por HCRF como por DML HF. En resumen, entre los RF, el HCRF parece ser el modelo superador en base a los niveles de *overfitting* y comportamiento de las métricas en la muestra de testeo.

En términos de la capacidad de los modelos de ordenar a la población según su reacción a la campaña, en general, la forma de las curvas Qini para KNN y OLS no es la deseada. Dependiendo la campaña y el *outcome*, se observan curvas horizontales o con poca pendiente. En el caso de OLS para 'visit' y 'spend' de la campaña masculina, incluso ocurre que el primer cuantil presenta un *uplift* promedio mayor que el último. En cambio, los modelos de RF presentan curvas monótonas crecientes con pendientes mayores, y cubriendo un mayor rango de valores en el eje y, incluso valores negativos, permitiendo mayor discriminación de los individuos.

Considerando todo lo anterior, parece conveniente elegir los modelos de 'visit' ya que lo desbalanceado de los otros *outcomes* genera modelos muy inestables y poco robustos, basándose en las métricas consideradas. En el caso de 'visit', las métricas y la capacidad de ordenamiento, globalmente sugieren que los RF son superadores, los cuales logran un *uplift* mayor al efecto promedio al seleccionar sólo un 10% de la población total en cada campaña, no sólo en el *outcome* modelado sino también en los otros *outcomes* de interés.

En resumen, en lo que respecta al objetivo de analizar ambas campañas, si bien los tres *outcomes* de interés indican consistentemente que la campaña de mercadería masculina es la más exitosa, los modelos de la campaña femenina parecen identificar de manera más eficiente a los mejores individuos. Es decir, en la campaña masculina es más marcada la necesidad de dirigir la campaña a una mayor proporción de clientes para aumentar el *uplift* y eso es una desventaja.

Por otro lado, en cuanto a la comparación de los distintos modelos cubiertos en esta tesis, aunque globalmente los RF parecen superiores, y se cumple la reducción del *overfitting* al considerar árboles honestos (Wager y Athey [19]), se esperaba encontrar una superioridad más marcada de los métodos de Machine Learning. Sin embargo, esto no es evidente al observar individualmente cada métrica, debido a ciertas características de los datos.

En primer lugar, la cantidad de variables disponibles es limitada. Aunque esto no es un inconveniente para el método de mínimos cuadrados, donde se consideró sólo un atributo predictor, ni para KNN que es un método que funciona bien incluyendo pocas variables, sí es una desventaja para los Random Forests, ya que no cuentan con esa herramienta para destacarse por sobre los otros modelos. Relacionado con este aspecto, el poder predictivo de las variables para estimar el efecto de tratamiento también mostró limitaciones. Por ejemplo, al considerar el método de MCO, en la mayoría de los casos el test de significatividad individual arrojó resultados

desfavorables. Además, utilizando el criterio de importancia de variables de la librería de Python EconML, se observó que la variable en primer lugar en cada caso de estudio era significativamente más importante que las siguientes.

En segundo lugar, aparecieron dificultades a la hora de modelar el *uplift* en la cantidad de compras ('conversion') y el monto de dichas compras ('spend') por lo desbalanceado de estas variables. Esto repercutió en que las métricas en general mostraron valores significativamente peores en la muestra de testeo que en la de entrenamiento.

Dadas estas dificultades encontradas en la estructura de los datos, otras futuras líneas de trabajo podrían evaluar si se puede mejorar la *performance* de los modelos de *uplift* en 'conversion' y 'spend', al considerar técnicas de *oversampling* o pesos. En cuanto a la limitación de las variables, se podría aprovechar las funciones de importancia de las variables, con las que cuentan las nuevas librerías diseñadas específicamente para estimar efectos de tratamiento, como las utilizadas en Python y "grf" de R. Esta herramienta puede ser de utilidad para elegir variables en métodos clásicos como OLS y KNN.

6 Apéndice I

Tabla 6-1: Frecuencias relativas de cada tratamiento según la variable 'recency'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
recency				
1	4.63%	4.64%	4.72%	13.99%
2	3.97%	3.95%	3.85%	11.78%
3	3.07%	3.02%	3.13%	9.22%
4	2.63%	2.69%	2.61%	7.93%
5	2.29%	2.37%	2.38%	7.05%
6	2.38%	2.39%	2.43%	7.20%
7	2.05%	2.20%	2.12%	6.37%
8	1.83%	1.82%	1.81%	5.46%
9	3.44%	3.32%	3.30%	10.06%
10	3.97%	3.88%	3.97%	11.82%
11	1.82%	1.80%	1.86%	5.47%
12	1.21%	1.20%	1.23%	3.64%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-2: Frecuencias relativas de cada tratamiento según la variable 'history_segment'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
history_segment				
1) 0−100	12.07%	11.89%	11.93%	35.89%
2) 100−200	7.33%	7.56%	7.39%	22.27%
3) 200−350	6.39%	6.32%	6.49%	19.20%
4) 350−500	3.28%	3.32%	3.42%	10.01%
5) 500−750	2.50%	2.58%	2.60%	7.67%
6) 750−1,000	1.01%	0.97%	0.93%	2.90%
7) $1,000 +	0.73%	0.65%	0.67%	2.04%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-3: Frecuencias relativas de cada tratamiento según la variable 'mens'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
mens				
0	14.95%	14.87%	15.07%	44.90%
1	18.34%	18.42%	18.34%	55.10%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-4: Frecuencias relativas de cada tratamiento según la variable 'womens'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
womens				
0	14.93%	15.06%	15.03%	45.03%
1	18.36%	18.23%	18.38%	54.97%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-5: Frecuencias relativas de cada tratamiento según la variable 'history_segment'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
zip_code				
Rural	5.07%	4.90%	4.97%	14.94%
Surburban	14.85%	15.04%	15.08%	44.96%
Urban	13.38%	13.35%	13.37%	40.10%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-6: Frecuencias relativas de cada tratamiento según la variable 'zip_code'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
newbie				
0	16.60%	16.58%	16.60%	49.78%
1	16.70%	16.71%	16.82%	50.22%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-7: Frecuencias relativas de cada tratamiento según la variable 'newbie'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
newbie				
0	16.60%	16.58%	16.60%	49.78%
1	16.70%	16.71%	16.82%	50.22%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-8: Frecuencias relativas de cada tratamiento según la variable 'Channel'

segment	Mens E-Mail	No E-Mail	Womens E-Mail	All
channel				
Multichannel	4.03%	4.07%	4.03%	12.13%
Phone	14.44%	14.57%	14.77%	43.78%
Web	14.83%	14.65%	14.62%	44.09%
All	33.29%	33.29%	33.42%	100.00%

Tabla 6-9: Lista final de Variables (incluyendo aquellas definidas por el modelador)

Recency	int64
recency_Q	object
history_segment	object
History	float64
Mens	int64
womens	int64
zip_code	object
Newbie	int64

channel	Object
segment	Category
visit	int64
conversion	int64
spend	float64
T_w	Int32
T_m	int32
e_score	float64
history_seg1	uint8
history_seg2	uint8
history_seg3	uint8
history_seg4	uint8
history_seg5	uint8
history_seg6	uint8
history_seg7	uint8
zip_code_Rural	uint8
zip_code_Surburban	uint8
zip_code_Urban	uint8
channel_Multichannel	uint8
channel_Phone	uint8
channel_Web	uint8

Tabla 6-10: VSI (Estabilidad entre Train y Test) – Campaña Femenina

recency	Train (%)	Test (%)	VSI
1	13.995%	14.159%	1.9e-05
2	11.802%	11.278%	0.00024
3	9.273%	9.064%	4.7e-05
4	7.996%	7.741%	8.3e-05
5	7.132%	7.097%	1.8e-06
6	7.232%	7.202%	1.2e-06
7	6.433%	6.675%	9e-05
8	5.422%	5.504%	1.2e-05
9	9.847%	10.235%	0.00015
10	11.785%	11.734%	2.2e-06
11	5.452%	5.598%	3.9e-05
12	3.631%	3.712%	1.8e-05
Total	100.000%	100.000%	0.0007

mens	Train (%)	Test (%)	VSI
0	44.923%	44.771%	5.1e-06
1	55.077%	55.229%	4.2e-06
Total	100.000%	100.000%	9.3e-06

womens	Train (%)	Test (%)	VSI
0	45.066%	45.298%	1.2e-05
1	54.934%	54.702%	9.8e-06
Total	100.000%	100.000%	2.2e-05

newbie	Train (%)	Test (%)	VSI
0	49.748%	49.701%	4.4e-07
1	50.252%	50.299%	4.4e-07
Total	100.000%	100.000%	8.8e-07

zip_code	Train (%)	Test (%)	VSI
Rural	14.812%	14.768%	1.4e-06
Surburban	45.099%	45.345%	1.3e-05
Urban	40.089%	39.888%	1e-05
Total	100.000%	100.000%	2.5e-05

channel	Train (%)	Test (%)	VSI
Multichannel	12.113%	12.273%	2.1e-05
Phone	43.863%	44.502%	9.2e-05
Web	44.024%	43.225%	0.00015
Total	100.000%	100.000%	0.00026

history_segment	Train (%)	Test (%)	VSI
1) 0–100	35.756%	35.531%	1.4e-05
2) 100–200	22.475%	22.099%	6.3e-05
3) 200–350	19.137%	19.475%	5.9e-05
4) 350–500	10.087%	10.153%	4.4e-06
5) 500–750	7.695%	8.034%	0.00015
6) 750–1,000	2.828%	2.916%	2.7e-05
7) $1,000 +	2.023%	1.792%	0.00028
Total	100.000%	100.000%	0.0006

Tabla 6-11: VSI (Estabilidad entre Train y Test) – Campaña Masculina

recency	Train (%)	Test (%)	VSI
1	13.910%	13.986%	4.1e-06
2	11.828%	12.226%	0.00013
3	9.226%	8.847%	0.00016
4	8.046%	7.779%	9e-05
5	7.005%	7.005%	2.4e-10
6	7.190%	7.005%	4.8e-05
7	6.424%	6.218%	6.7e-05
8	5.500%	5.421%	1.2e-05
9	10.235%	9.867%	0.00013
10	11.608%	12.496%	0.00065
11	5.480%	5.256%	9.3e-05
12	3.549%	3.895%	0.00032
Total	100.000%	100.000%	0.0017

mens	Train (%)	Test (%)	VSI
0	44.849%	44.562%	1.8e-05
1	55.151%	55.438%	1.5e-05
Total	100.000%	100.000%	3.3e-05

womens			
0	45.057%	45.008%	5.5e-07
1	54.943%	54.992%	4.5e-07
Total	100.000%	100.000%	9.9e-07

newbie	Train (%)	Test (%)	VSI
0	49.833%	49.795%	2.9e-07
1	50.167%	50.205%	2.9e-07
Total	100.000%	100.000%	5.8e-07

zip_code	Train (%)	Test (%)	VSI
Rural	15.057%	14.654%	0.00011
Surburban	44.667%	45.747%	0.00026
Urban	40.276%	39.599%	0.00011
Total	100.000%	100.000%	0.00048

channel	Train (%)	Test (%)	VSI
Multichannel	12.144%	12.237%	7.1e-06
Phone	43.608%	43.424%	7.8e-06
Web	44.248%	44.339%	1.9e-06
Total	100.000%	100.000%	1.7e-05

history_segment	Train (%)	Test (%)	VSI
1) 0–100	36.113%	35.492%	0.00011
2) 100–200	22.253%	22.774%	0.00012
3) 200–350	19.287%	18.292%	0.00053
4) 350–500	9.824%	10.231%	0.00017
5) 500–750	7.580%	7.802%	6.4e-05
6) 750–1,000	2.922%	3.168%	0.0002
7) $1,000 +	2.021%	2.241%	0.00023
Total	100.000%	100.000%	0.0014

7 Apéndice II

7.1. Apéndice II -1

Observación: En los árboles de regresión, encontrar el corte que minimice el error cuadrático entre $\hat{\mu}(X)$ e Y es equivalente a maximizar la suma de $\hat{\mu}(x_i)^2$ sobre la muestra de entrenamiento S (donde $\hat{\mu}$ en cada hoja es el promedio de la variable de respuesta Y).

Demostración:

Basta probar que vale la siguiente igualdad, dado que en el término de la derecha, la primera suma no depende de la estructura del árbol:

$$\sum_{i \in S} (\hat{\mu}(x_i) - Y_i)^2 = \sum_{i \in S} (Y_i)^2 - \sum_{i \in S} \left(\hat{\mu}(x_i)\right)^2$$

Se notará por $l = 1, \dots, L$ a las hojas del árbol, y se utilizará el hecho de que para cada $i \in l$ por definición se tiene que $\hat{\mu}(x_{i_l}) = \frac{1}{\#\{i \in l\}} \sum_{\{i \in L\}} Y_i$, y que este valor es constante dentro de cada hoja.

Entonces:

$$\sum_{i \in S} (\hat{\mu}(x_i) - Y_i)^2 = \sum_{i \in S} \left(\hat{\mu}(x_i)\right)^2 + \sum_{i \in S} (Y_i)^2 - 2 \underbrace{\sum_{i \in S} \hat{\mu}(x_i) \cdot Y_i}_{(*)}$$

$$\begin{aligned} (*) &= \sum_{l=1}^{L} \sum_{i \in l} \hat{\mu}(x_i) \cdot Y_i \\ &= \sum_{l=1}^{L} \hat{\mu}(x_{i_l}) \sum_{i \in l} Y_i \\ &= \sum_{l=1}^{L} \hat{\mu}(x_{i_l}) \cdot \#\{i \in l\}\, \hat{\mu}(x_{i_l}) \\ &= \sum_{l=1}^{L} \sum_{i \in l} \left(\hat{\mu}(x_i)\right)^2 \end{aligned}$$

□

7.2. Apéndice II -2: Resultados Regresión Lineal

Tabla 7-1: MCO con un atributo para la campaña de mercadería femenina - total

Variable	T_w	Param T	t stat T	p val T	Param Inter	t stat Inter	p val Inter	R^2	R^2 adj
recency	Visit	0.041066	6.616109	0.0000	0.000746	0.809685	0.4181	0.009852	0.009783
	Conversion	0.002654	1.677270	0.0935	0.000081	0.346073	0.7293	0.000992	0.000922
	Spend	0.517654	2.064027	0.0390	-0.015998	-0.430020	0.6672	0.000507	0.000437
history	Visit	0.043118	9.674008	0.0000	0.000008	0.643573	0.5199	0.008552	0.008482
	Conversion	0.002220	1.954646	0.0506	0.000004	1.120079	0.2627	0.001039	0.000969
	Spend	0.251889	1.399695	0.1616	0.000707	1.376989	0.1685	0.000671	0.000601
mens	Visit	0.073985	15.333562	0.0000	-0.052232	-8.036071	0.0000	0.006184	0.006114
	Conversion	0.005115	4.163992	0.0000	-0.003642	-2.201005	0.0277	0.000469	0.000399
	Spend	0.631648	3.245557	0.0012	-0.375232	-1.431276	0.1524	0.000311	0.000241
womens	Visit	0.011078	2.306945	0.0211	0.062036	9.570528	0.0000	0.010696	0.010627
	Conversion	0.000633	0.516458	0.6055	0.004503	2.723336	0.0065	0.000752	0.000682
	Spend	0.278354	1.433761	0.1516	0.265859	1.014529	0.3103	0.000276	0.000206
newbie	Visit	0.040123	8.770400	0.0000	0.010295	1.595362	0.1106	0.010130	0.010060
	Conversion	0.000932	0.798585	0.4245	0.004340	2.637227	0.0084	0.000611	0.000541
	Spend	0.113957	0.616396	0.5376	0.618331	2.371124	0.0177	0.000468	0.000398

Tabla 7-2: MCO con un atributo para la campaña de mercadería femenina - train

Variable	T_w	Param T	t stat T	p val T	Param Inter	t stat Inter	p val Inter	R^2	R^2 adj
recency	Visit	0.040981	5.929349	0.0000	0.001008	0.982186	0.3260	0.010220	0.010133
	Conversion	0.002894	1.649412	0.0991	0.000020	0.076001	0.9394	0.000988	0.000900
	Spend	0.405986	1.553800	0.1202	-0.010857	-0.279737	0.7797	0.000535	0.000447
history	Visit	0.045200	9.120107	0.0000	0.000005	0.328902	0.7422	0.008641	0.008554
	Conversion	0.001999	1.590652	0.1117	0.000004	1.115620	0.2646	0.001201	0.001113
	Spend	0.269941	1.441914	0.1493	0.000286	0.536725	0.5915	0.000520	0.000432
mens	Visit	0.071868	13.365600	0.0000	-0.046177	-6.373276	0.0000	0.006161	0.006073
	Conversion	0.004297	3.153761	0.0016	-0.002397	-1.305798	0.1916	0.000365	0.000277
	Spend	0.444164	2.189152	0.0286	-0.188730	-0.690333	0.4900	0.000215	0.000128
womens	Visit	0.015513	2.896567	0.0038	0.056487	7.816991	0.0000	0.010882	0.010795
	Conversion	0.000870	0.639618	0.5224	0.003847	2.096654	0.0360	0.000708	0.000620
	Spend	0.289558	1.429393	0.1529	0.092468	0.338321	0.7351	0.000189	0.000101
newbie	Visit	0.041391	8.114933	0.0000	0.010048	1.396478	0.1626	0.009684	0.009597
	Conversion	0.000933	0.721030	0.4709	0.004066	2.226331	0.0260	0.000505	0.000418
	Spend	0.045676	0.236922	0.8127	0.586271	2.155742	0.0311	0.000379	0.000292

Tabla 7-3: MCO con un atributo para la campaña de mercadería masculina - total

		Param T	t stat T	p val T	Param Inter	t stat Inter	p val Inter	R^2	R^2 adj
Variable	T_m								
recency	Visit	0.085378	13.149318	0.0000	-0.001492	-1.549581	0.1212	0.018395	0.018326
	Conversion	0.007271	4.105404	0.0000	-0.000078	-0.297190	0.7663	0.001949	0.001879
	Spend	0.973611	3.484409	0.0005	-0.035075	-0.846572	0.3972	0.000950	0.000880
history	Visit	0.070931	15.275422	0.0000	0.000023	1.720616	0.0853	0.016323	0.016253
	Conversion	0.004842	3.827343	0.0001	0.000008	2.236812	0.0253	0.002277	0.002207
	Spend	0.512444	2.567431	0.0102	0.001054	1.862280	0.0626	0.001174	0.001103
mens	Visit	0.070580	13.959239	0.0000	0.010986	1.614444	0.1064	0.013100	0.013030
	Conversion	0.005722	4.158707	0.0000	0.001970	1.063915	0.2874	0.001419	0.001349
	Spend	0.615731	2.837375	0.0046	0.280907	0.961816	0.3361	0.000887	0.000817
womens	Visit	0.069155	13.713154	0.0000	0.013403	1.970191	0.0488	0.012626	0.012557
	Conversion	0.006010	4.380004	0.0000	0.001440	0.777800	0.4367	0.001329	0.001259
	Spend	0.682735	3.154745	0.0016	0.158476	0.542837	0.5872	0.000666	0.000595
newbie	Visit	0.077833	16.278337	0.0000	-0.002528	-0.374489	0.7080	0.018251	0.018182
	Conversion	0.006207	4.758300	0.0000	0.001189	0.645758	0.5184	0.001608	0.001538
	Spend	0.577967	2.808985	0.0050	0.382053	1.315264	0.1884	0.000853	0.000783

Tabla 7-4: MCO con un atributo para la campaña de mercadería masculina – train

		Param T	t stat T	p val T	Param Inter	t stat Inter	p val Inter	R^2	R^2 adj
Variable	T_m								
recency	Visit	0.089331	12.316483	0.0000	-0.001936	-1.797922	0.0722	0.018800	0.018713
	Conversion	0.007082	3.552438	0.0004	0.000004	0.014204	0.9887	0.001818	0.001730
	Spend	0.858069	2.739214	0.0062	-0.019050	-0.409652	0.6821	0.000758	0.000670
history	Visit	0.071979	13.885711	0.0000	0.000025	1.693063	0.0905	0.017237	0.017151
	Conversion	0.005530	3.884961	0.0001	0.000006	1.598982	0.1098	0.002349	0.002261
	Spend	0.514608	2.300848	0.0214	0.000965	1.514111	0.1300	0.001143	0.001055
mens	Visit	0.075837	13.445413	0.0000	0.004196	0.552451	0.5806	0.013660	0.013573
	Conversion	0.007109	4.596538	0.0000	-0.000010	-0.004994	0.9960	0.001455	0.001367
	Spend	0.802830	3.304263	0.0010	-0.097322	-0.297468	0.7661	0.000776	0.000688
womens	Visit	0.069302	12.311720	0.0000	0.015694	2.066654	0.0388	0.013089	0.013002
	Conversion	0.005844	3.787650	0.0002	0.002259	1.084944	0.2780	0.001479	0.001391
	Spend	0.463872	1.913525	0.0557	0.515196	1.575306	0.1152	0.000699	0.000611
newbie	Visit	0.079376	14.867989	0.0000	-0.002825	-0.374855	0.7078	0.018104	0.018017
	Conversion	0.007146	4.870896	0.0000	-0.000109	-0.052844	0.9579	0.001628	0.001540
	Spend	0.574651	2.493105	0.0127	0.343528	1.055624	0.2911	0.000772	0.000684

8 Apéndice III: Comparación – Campaña Femenina

Campaña Femenina – visit

Tabla 8-1: *Uplift* en 'visit' en función de la variable 'womens'

'womens'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
0	4,80%	7,19%	7,19%	7,17%	7,15%
1	4,49%	2,57%	2,60%	2,58%	2,59%
Muestra de Testeo					
0	4,81%	7,19%	7,28%	7,14%	7,16%
1	4,49%	2,57%	2,74%	2,61%	2,67%

Tabla 8-2: *Uplift* en 'visit' en función de la variable 'history'

'history'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	5,04%	4,88%	4,98%	5,03%	5,02%
2) 200	4,46%	4,85%	4,50%	4,40%	4,43%
3) 350	3,37%	4,45%	3,55%	3,57%	3,48%
4) 500	3,32%	4,43%	3,42%	3,49%	3,55%
5) 750	6,88%	4,09%	7,07%	6,59%	6,73%
6) 1,000	6,14%	4,07%	5,95%	6,73%	6,65%
7) $1,000 +	6,72%	3,97%	6,51%	5,80%	6,01%
Muestra de Testeo					
1) 100	5,01%	4,86%	4,91%	4,93%	4,93%
2) 200	4,41%	4,88%	4,58%	4,45%	4,46%
3) 350	3,50%	4,36%	3,80%	3,53%	3,49%
4) 500	3,46%	4,42%	4,39%	3,72%	4,07%
5) 750	6,82%	4,33%	7,10%	6,80%	6,93%
6) 1,000	6,02%	3,89%	5,37%	6,44%	6,33%
7) $1,000 +	6,80%	3,93%	5,93%	5,66%	6,02%

Tabla 8-3: *Uplift* en 'visit' en función de la variable 'mens'

'mens'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
0	4,80%	7,19%	7,19%	7,17%	7,15%
1	4,49%	2,57%	2,60%	2,58%	2,59%
Muestra de Testeo					
0	4,81%	7,19%	7,28%	7,14%	7,16%
1	4,49%	2,57%	2,74%	2,61%	2,67%

Tabla 8-4: *Uplift* en 'visit' en función de la variable 'recency'

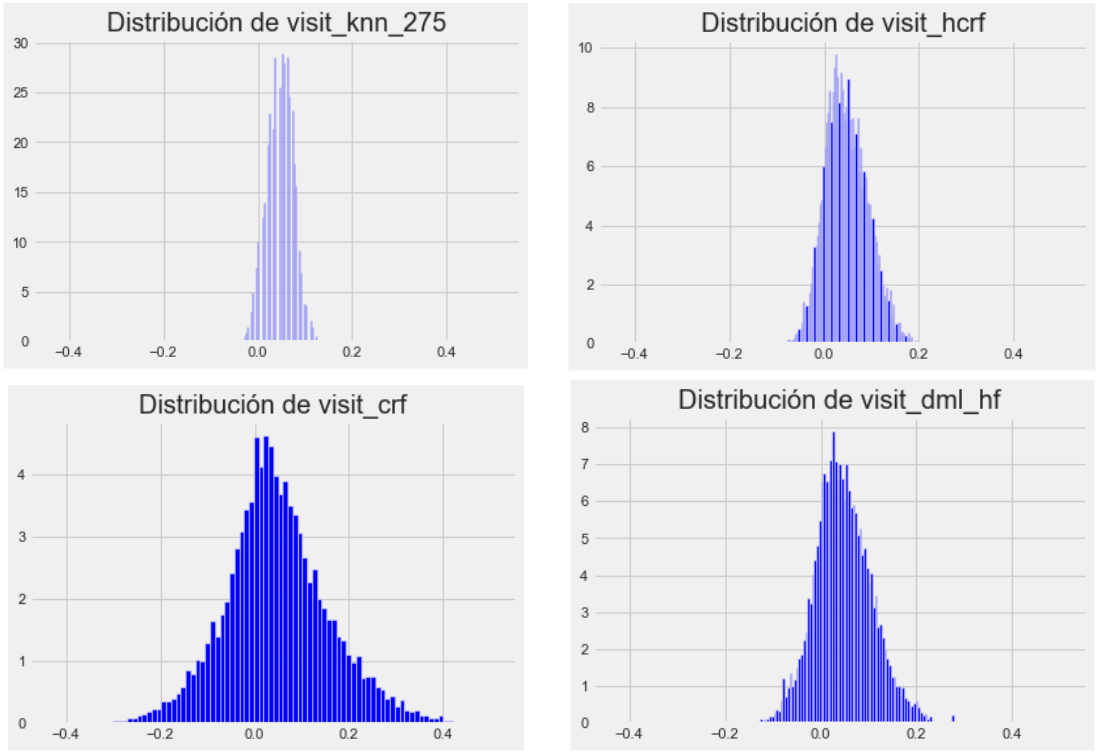
'recency'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					

1	4,45%	4,54%	4,69%	4,47%	4,42%
2	4,33%	4,50%	4,09%	4,14%	4,08%
3	4,44%	4,68%	4,14%	4,47%	4,29%
4	4,71%	4,68%	4,24%	4,32%	4,34%
5	4,63%	4,67%	4,06%	4,38%	4,49%
6	4,69%	4,59%	4,35%	4,63%	4,50%
7	4,90%	4,63%	5,55%	5,24%	5,37%
8	5,23%	4,60%	5,43%	5,08%	5,19%
9	4,88%	4,71%	5,79%	5,26%	5,40%
10	4,58%	4,75%	4,07%	4,53%	4,50%
11	4,51%	4,74%	4,47%	4,75%	4,79%
12	4,64%	4,82%	6,89%	5,35%	5,56%
Muestra de Testeo					
1	4,54%	4,57%	4,61%	4,51%	4,43%
2	4,41%	4,55%	3,87%	4,08%	4,04%
3	4,33%	4,56%	4,04%	4,28%	4,10%
4	4,92%	4,65%	4,62%	4,55%	4,69%
5	4,66%	4,58%	4,25%	4,31%	4,43%
6	4,73%	4,66%	4,43%	4,84%	4,78%
7	4,93%	4,53%	5,39%	4,96%	5,10%
8	5,15%	4,56%	5,22%	4,85%	4,97%
9	4,80%	4,77%	5,96%	5,24%	5,42%
10	4,43%	4,73%	4,48%	4,49%	4,57%
11	4,71%	4,78%	5,23%	5,08%	5,18%
12	4,32%	4,77%	7,23%	5,28%	5,58%

Tabla 8-5: *Uplift* en 'visit' en función de la variable 'zip_code'

'zip_code'	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
Rural	4,55%	4,63%	3,33%	3,96%	3,97%
Surburban	4,64%	4,67%	5,22%	4,91%	4,91%
Urban	4,64%	4,62%	4,52%	4,58%	4,59%
Muestra de Testeo					
Rural	4,55%	4,58%	3,40%	3,93%	4,06%
Surburban	4,67%	4,66%	5,26%	4,89%	4,87%
Urban	4,62%	4,64%	4,74%	4,61%	4,69%

Figura 8-1: Distribución de las predicciones continuas del CATE en 'visit' en la población total (WOMEN_campaign)



Campaña Femenina – conversion

Tabla 8-6: *Uplift* en 'conversion' en función de la variable 'history'

'history_segment'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	0,37%	-0,77%	0,43%	0,40%	0,38%
2) 200	0,36%	-0,78%	0,33%	0,27%	0,35%
3) 350	-0,12%	-0,99%	-0,05%	0,01%	-0,07%
4) 500	-0,22%	-1,00%	-0,19%	-0,07%	-0,22%
5) 750	0,81%	-1,18%	0,84%	0,81%	0,81%
6) 1,000	1,46%	-1,19%	1,41%	1,24%	1,46%
7) $1,000 +	1,72%	-1,24%	1,00%	1,03%	1,34%
Muestra de Testeo					
1) 100	0,37%	-0,78%	0,45%	0,40%	0,38%
2) 200	0,34%	-0,77%	0,28%	0,23%	0,30%
3) 350	-0,07%	-1,04%	-0,02%	0,03%	-0,04%
4) 500	-0,25%	-1,01%	-0,16%	0,01%	-0,17%
5) 750	0,79%	-1,06%	0,81%	0,81%	0,82%
6) 1,000	1,46%	-1,28%	1,67%	1,28%	1,59%
7) $1,000 +	1,64%	-1,26%	0,63%	0,91%	1,20%

Tabla 8-7: *Uplift* en 'conversion' en función de la variable 'recency'

'recency	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					

1	0,32%	-0,94%	0,44%	0,45%	0,46%
2	0,28%	-0,97%	0,32%	0,27%	0,30%
3	0,32%	-0,87%	0,03%	0,06%	0,10%
4	0,37%	-0,87%	0,39%	0,30%	0,31%
5	0,23%	-0,88%	0,45%	0,41%	0,40%
6	0,22%	-0,92%	0,20%	0,23%	0,20%
7	0,26%	-0,90%	0,11%	0,20%	0,21%
8	0,27%	-0,91%	-0,01%	0,25%	0,13%
9	0,35%	-0,86%	0,52%	0,40%	0,42%
10	0,36%	-0,83%	0,41%	0,47%	0,40%
11	0,33%	-0,84%	0,48%	0,31%	0,33%
12	0,33%	-0,80%	0,32%	0,28%	0,22%
Muestra de Testeo					
1	0,34%	-0,93%	0,54%	0,47%	0,49%
2	0,34%	-0,94%	0,36%	0,33%	0,34%
3	0,29%	-0,93%	0,11%	0,07%	0,16%
4	0,31%	-0,89%	0,37%	0,31%	0,33%
5	0,19%	-0,92%	0,41%	0,35%	0,34%
6	0,23%	-0,88%	0,15%	0,23%	0,16%
7	0,29%	-0,95%	0,10%	0,22%	0,21%
8	0,21%	-0,93%	-0,04%	0,18%	0,08%
9	0,29%	-0,82%	0,50%	0,39%	0,39%
10	0,35%	-0,85%	0,36%	0,46%	0,39%
11	0,40%	-0,82%	0,49%	0,36%	0,33%
12	0,33%	-0,83%	0,14%	0,19%	0,12%

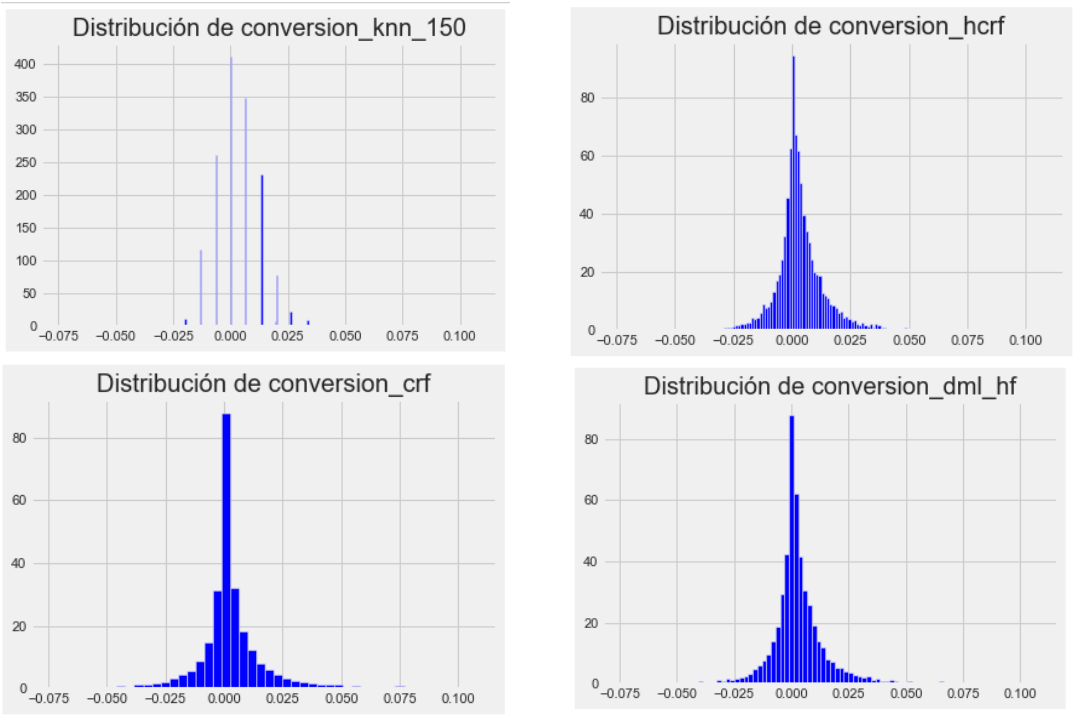
Tabla 8-8: *Uplift* en 'conversion' en función de la variable 'womens'

'womens	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
0	0,28%	-1,97%	0,13%	0,14%	0,18%
1	0,33%	-0,01%	0,48%	0,47%	0,42%
Muestra de Testeo					
0	0,30%	-1,97%	0,13%	0,11%	0,16%
1	0,31%	-0,01%	0,49%	0,49%	0,43%

Tabla 8-9: *Uplift* en 'conversion' en función de la variable 'mens'

'mens	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
0	0,31%	0,43%	0,46%	0,44%	0,41%
1	0,30%	-1,97%	0,22%	0,22%	0,23%
Muestra de Testeo					
0	0,30%	0,43%	0,50%	0,48%	0,43%
1	0,31%	-1,97%	0,18%	0,19%	0,21%

Figura 8-2: Distribución de las predicciones continuas del CATE en 'conversion' en la población total (WOMEN_campaign)



Campaña Femenina – spend

Tabla 8-10: *Uplift* en 'spend' en función de la variable 'history'

'history_segment'	knn_150	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	58,87%	34,98%	62,19%	51,08%	54,23%
2) 200	44,00%	34,88%	40,82%	39,51%	42,07%
3) 350	-28,01%	33,24%	-21,86%	-7,78%	-18,32%
4) 500	-31,05%	33,14%	-32,87%	-17,61%	-26,50%
5) 750	67,66%	31,76%	76,96%	65,44%	69,70%
6) 1,000	131,95%	31,68%	110,91%	123,26%	133,41%
7) $1,000 +	218,34%	31,25%	158,23%	154,69%	164,21%
Muestra de Testeo					
1) 100	55,96%	34,92%	54,22%	45,62%	46,51%
2) 200	41,10%	34,98%	35,69%	32,27%	35,33%
3) 350	-25,49%	32,87%	-16,40%	-7,76%	-16,68%
4) 500	-34,48%	33,12%	-36,30%	-19,77%	-26,51%
5) 750	70,29%	32,72%	74,96%	68,60%	71,05%
6) 1,000	121,56%	30,92%	120,19%	125,81%	139,55%
7) $1,000 +	199,83%	31,09%	85,32%	161,84%	171,60%

Tabla 8-11: *Uplift* en 'spend' en función de la variable 'recency'

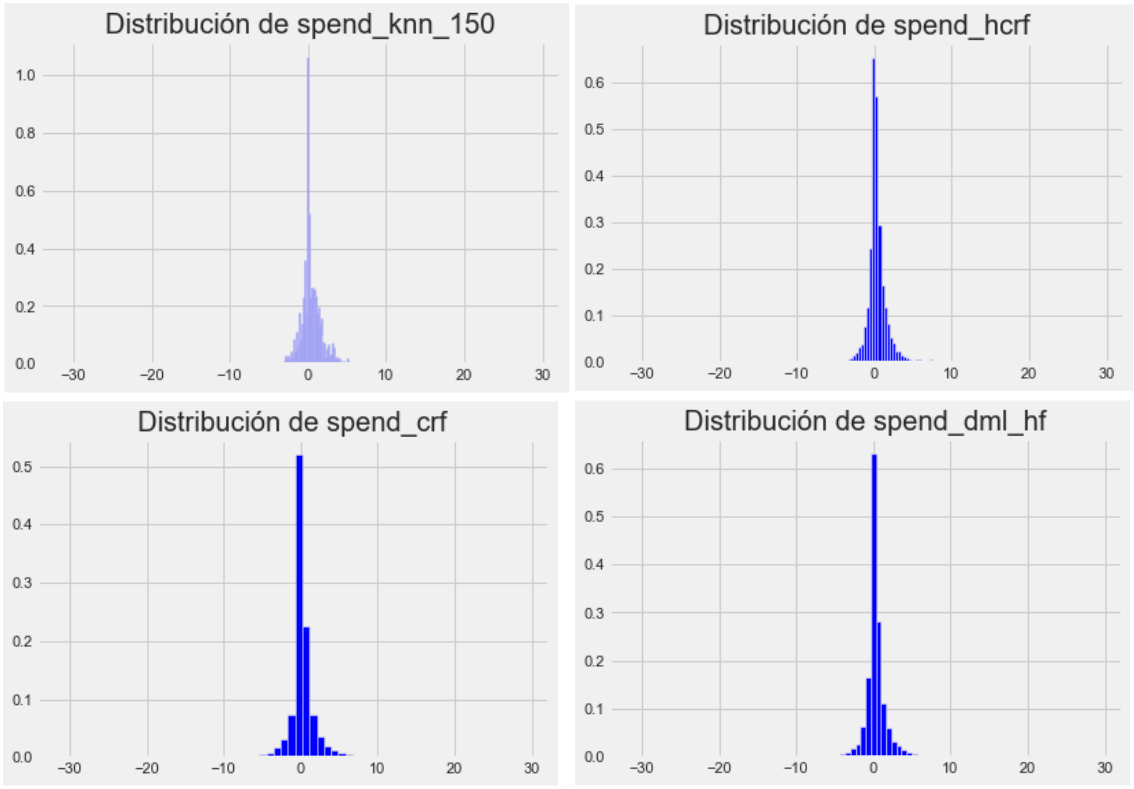
'recency'	knn_150	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					

1	39,21%	33,62%	41,36%	52,15%	44,49%
2	38,06%	33,43%	34,71%	30,41%	23,11%
3	41,59%	34,15%	-10,26%	10,14%	14,41%
4	66,27%	34,16%	108,25%	62,82%	77,08%
5	47,05%	34,11%	72,34%	56,85%	61,95%
6	28,35%	33,82%	35,89%	36,23%	38,85%
7	13,84%	33,96%	-1,93%	11,22%	6,97%
8	13,70%	33,84%	-9,88%	24,06%	8,56%
9	26,47%	34,29%	35,31%	29,98%	33,10%
10	36,38%	34,46%	34,69%	33,57%	36,54%
11	32,56%	34,40%	61,74%	41,42%	47,33%
12	26,41%	34,75%	15,34%	22,41%	10,61%
Muestra de Testeo					
1	37,89%	33,74%	38,60%	47,88%	43,06%
2	47,83%	33,64%	36,34%	32,46%	29,72%
3	38,96%	33,69%	-5,93%	13,54%	26,82%
4	52,45%	34,05%	98,72%	57,45%	65,13%
5	31,15%	33,77%	57,90%	43,23%	46,77%
6	31,98%	34,07%	29,48%	30,62%	39,82%
7	18,68%	33,56%	3,84%	12,57%	6,48%
8	7,89%	33,69%	-14,02%	13,87%	-11,23%
9	19,76%	34,55%	34,39%	25,73%	28,16%
10	31,25%	34,38%	17,95%	27,33%	24,09%
11	35,97%	34,59%	58,95%	43,68%	44,14%
12	27,57%	34,53%	0,50%	16,83%	3,80%

Tabla 8-12: *Uplift* en 'spend' en función de la variable 'newbie'

'newbie'	**knn_150**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
0	19,37%	34,27%	9,96%	10,36%	11,02%
1	52,07%	33,78%	62,12%	60,45%	58,99%
Muestra de Testeo					
0	16,29%	34,02%	5,53%	2,99%	3,95%
1	50,05%	33,97%	56,82%	60,47%	58,21%

Figura 8-3: Distribución de las predicciones continuas del CATE en 'spend' en la población total (WOMEN_campaign)



9 Apéndice IV: Comparación – Campaña Masculina

Campaña Masculina – visit

Tabla 9-1: *Uplift* en 'visit' en función de la variable 'history'

'history_segment'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	7,38%	7,33%	7,41%	7,33%	7,26%
2) 200	6,59%	7,56%	6,62%	6,58%	6,43%
3) 350	8,34%	7,87%	8,34%	8,30%	8,13%
4) 500	9,10%	8,24%	9,08%	8,86%	9,01%
5) 750	9,83%	8,71%	10,10%	9,93%	9,93%
6) 1,000	6,72%	9,34%	6,19%	7,37%	6,94%
7) $1,000 +	10,12%	10,45%	10,34%	9,39%	9,06%
Muestra de Testeo					
1) 100	7,37%	7,33%	7,21%	7,24%	7,12%
2) 200	6,51%	7,56%	6,30%	6,47%	6,24%
3) 350	8,47%	7,86%	8,57%	8,41%	8,27%
4) 500	9,00%	8,24%	8,45%	8,76%	8,93%
5) 750	9,95%	8,72%	10,43%	10,06%	10,11%
6) 1,000	6,51%	9,33%	6,38%	7,28%	6,79%
7) $1,000 +	10,16%	10,42%	7,99%	9,31%	8,52%

Tabla 9-2: *Uplift* en 'visit' en función de la variable 'recency'

'recency'	knn_275	Ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1	8,40%	8,09%	8,72%	8,44%	8,38%
2	8,19%	7,98%	8,30%	8,26%	8,07%
3	8,23%	7,87%	9,33%	8,53%	8,49%
4	8,21%	7,83%	8,47%	8,18%	7,96%
5	7,86%	7,78%	7,44%	7,73%	7,49%
6	7,38%	7,74%	7,96%	7,44%	7,34%
7	7,11%	7,72%	5,97%	6,86%	6,52%
8	7,38%	7,67%	6,67%	6,99%	6,93%
9	7,48%	7,67%	8,05%	7,55%	7,63%
10	7,47%	7,63%	6,49%	7,07%	7,07%
11	7,20%	7,62%	6,46%	7,05%	6,96%
12	7,12%	7,58%	8,13%	7,51%	7,49%
Muestra de Testeo					

1	8,53%	8,11%	8,63%	8,57%	8,44%
2	8,31%	7,96%	8,08%	8,21%	7,96%
3	7,95%	7,94%	8,72%	8,32%	8,13%
4	8,27%	7,83%	8,57%	8,33%	8,09%
5	7,75%	7,81%	7,22%	7,65%	7,37%
6	7,56%	7,77%	8,09%	7,47%	7,30%
7	7,03%	7,69%	6,10%	6,78%	6,57%
8	7,29%	7,72%	5,50%	6,76%	6,50%
9	7,42%	7,64%	7,88%	7,41%	7,42%
10	7,43%	7,65%	6,62%	7,07%	7,17%
11	7,31%	7,65%	6,17%	6,93%	6,91%
12	7,09%	7,57%	7,87%	7,46%	7,36%

Tabla 9-3: *Uplift* en 'visit' en función de la variable 'womens'

'womens'	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
0	7,69%	7,72%	7,04%	7,32%	7,22%
1	7,85%	7,87%	8,43%	8,09%	7,99%
Muestra de Testeo					
0	7,68%	7,72%	6,89%	7,28%	7,16%
1	7,86%	7,89%	8,21%	8,05%	7,91%

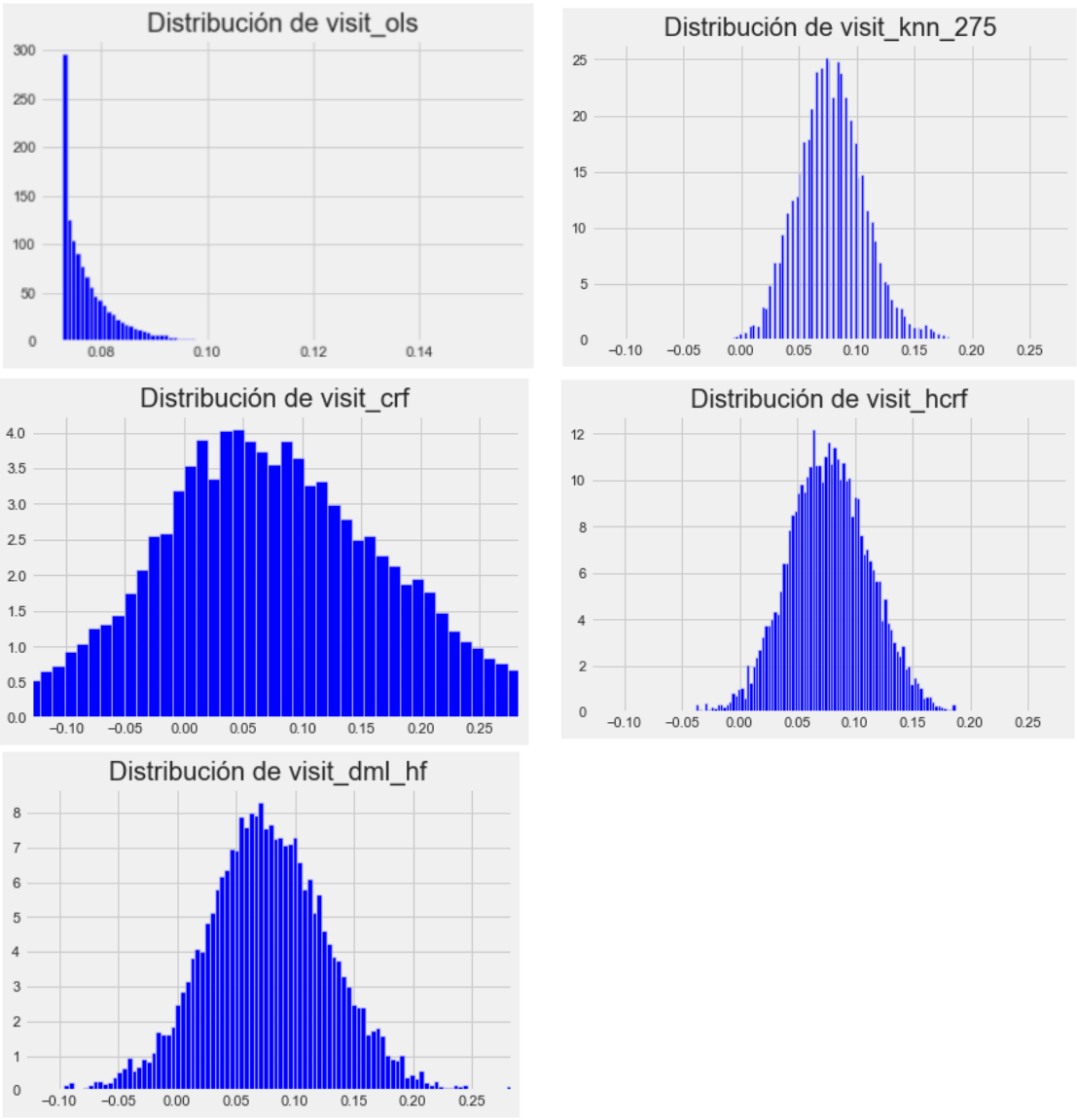
Tabla 9-4: *Uplift* en 'visit' en función de la variable 'zip_code'

'zip_code'	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
Rural	7,73%	7,81%	7,60%	7,67%	7,51%
Surburban	7,81%	7,80%	7,79%	7,71%	7,65%
Urban	7,77%	7,80%	7,89%	7,80%	7,69%
Muestra de Testeo					
Rural	7,66%	7,82%	6,83%	7,48%	7,22%
Surburban	7,76%	7,81%	7,48%	7,60%	7,51%
Urban	7,85%	7,82%	8,07%	7,90%	7,77%

Tabla 9-5: *Uplift* en 'visit' en función de la variable 'mens'

''	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
0	7,66%	7,72%	7,65%	7,67%	7,54%
1	7,88%	7,87%	7,93%	7,80%	7,73%
Muestra de Testeo					
0	7,66%	7,73%	7,55%	7,62%	7,45%
1	7,88%	7,88%	7,67%	7,77%	7,66%

Figura 9-1: Distribución de las predicciones continuas del CATE en 'visit' en la población total (MEN_campaign)



Campaña Masculina – conversion

Tabla 9-6: *Uplift* en 'conversion' en función de la variable 'history'

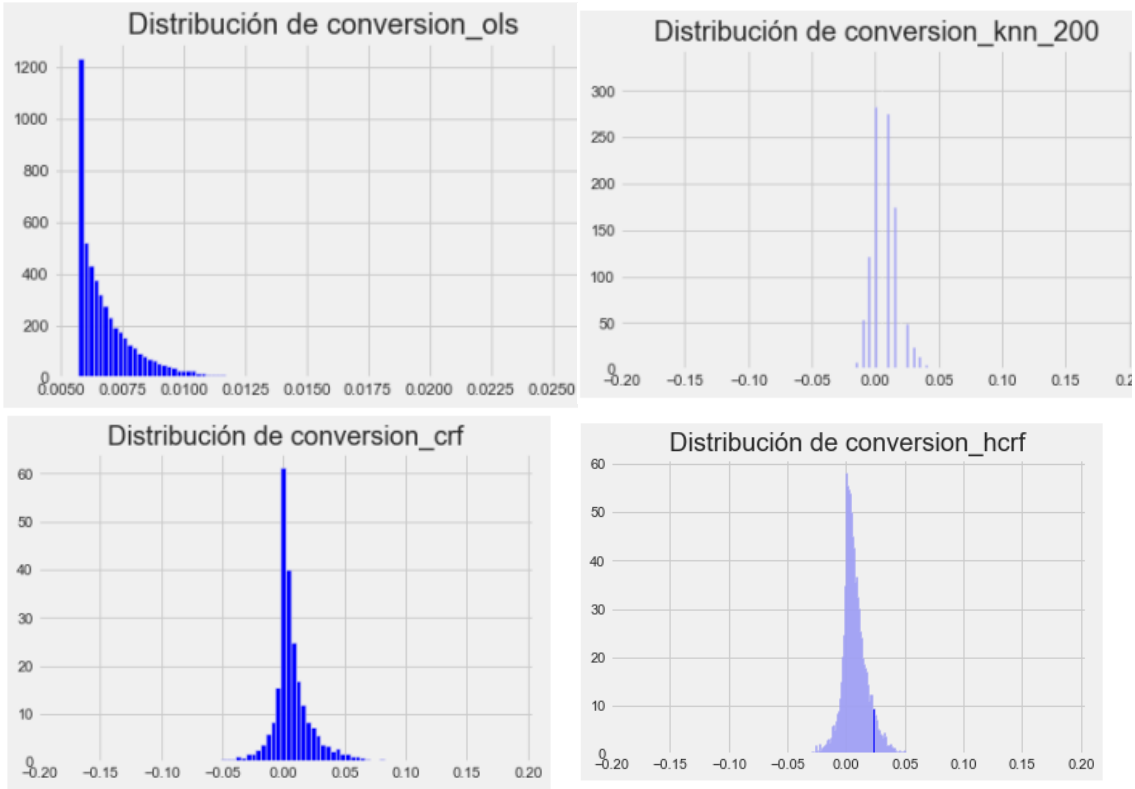
'history_segment'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	0,54%	0,58%	0,54%	0,55%	0,55%
2) 200	0,57%	0,64%	0,53%	0,53%	0,54%
3) 350	0,93%	0,71%	0,95%	0,97%	1,03%
4) 500	1,06%	0,80%	1,05%	0,99%	1,06%
5) 750	0,44%	0,91%	0,45%	0,62%	0,42%
6) 1,000	1,73%	1,07%	1,78%	1,25%	1,28%
7) $1,000 +	1,23%	1,33%	1,32%	1,20%	1,04%
Muestra de Testeo					

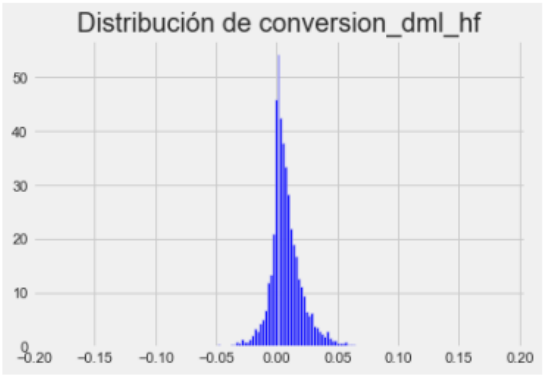
1) 100	0,55%	0,58%	0,57%	0,54%	0,55%
2) 200	0,57%	0,64%	0,60%	0,52%	0,55%
3) 350	0,94%	0,71%	0,92%	0,96%	1,03%
4) 500	1,03%	0,80%	1,05%	1,06%	1,05%
5) 750	0,42%	0,92%	0,51%	0,64%	0,48%
6) 1,000	1,83%	1,06%	2,03%	1,29%	1,47%
7) $1,000 +	1,19%	1,33%	1,41%	1,17%	1,00%

Tabla 9-7: *Uplift* en 'conversion' en función de la variable 'zip_code'

'zip_code'	**knn_275**	**ols**	**Crf**	**hcrf**	**dml_hf**
Muestra de Entrenamiento					
Rural	0,70%	0,70%	0,51%	0,52%	0,48%
Surburban	0,71%	0,70%	0,79%	0,78%	0,79%
Urban	0,71%	0,70%	0,70%	0,70%	0,72%
Muestra de Testeo					
Rural	0,70%	0,70%	0,52%	0,49%	0,46%
Surburban	0,72%	0,70%	0,83%	0,77%	0,78%
Urban	0,73%	0,70%	0,75%	0,72%	0,75%

Figura 9-2: Distribución de las predicciones continuas del CATE en 'conversion' en la población total (MEN_campaign)





Campaña Maculina – spend

Tabla 9-8: *Uplift* en 'spend' en función de la variable 'history'

'history'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1) 100	50,48%	56,46%	57,51%	64,34%	64,12%
2) 200	82,04%	65,58%	88,63%	79,73%	86,76%
3) 350	49,77%	77,25%	65,74%	76,82%	73,25%
4) 500	146,68%	91,68%	142,45%	139,79%	150,61%
5) 750	59,89%	109,67%	48,00%	76,26%	65,34%
6) 1,000	178,65%	133,99%	134,65%	128,30%	155,40%
7) $1,000 +	112,89%	176,96%	150,93%	102,25%	134,04%
Muestra de Testeo					
1) 100	51,53%	56,38%	61,70%	62,20%	62,55%
2) 200	79,23%	65,57%	91,85%	77,61%	85,20%
3) 350	50,02%	77,17%	65,20%	77,82%	79,01%
4) 500	151,16%	91,79%	140,46%	148,42%	150,86%
5) 750	55,33%	110,23%	52,14%	82,76%	68,52%
6) 1,000	190,10%	133,74%	135,46%	129,43%	144,95%
7) $1,000 +	122,74%	175,78%	271,49%	128,38%	155,23%

Tabla 9-9: *Uplift* en 'spend' en función de la variable 'recency'

'recency'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
1	88,33%	85,77%	106,01%	88,30%	91,15%
2	81,05%	81,55%	55,22%	82,94%	100,63%
3	74,61%	77,30%	114,47%	127,85%	113,55%
4	76,07%	75,89%	75,29%	79,18%	73,26%
5	65,43%	73,75%	54,73%	68,65%	65,97%
6	63,60%	72,32%	58,20%	53,20%	66,72%
7	48,92%	71,72%	82,46%	70,36%	77,09%
8	53,17%	69,78%	-32,51%	19,14%	5,39%
9	72,30%	69,57%	74,01%	76,82%	83,35%
10	71,01%	68,00%	87,28%	90,91%	95,28%
11	66,64%	67,62%	62,46%	79,60%	62,49%
12	88,46%	66,17%	203,89%	109,32%	139,57%
Muestra de Testeo					

1	94,75%	86,72%	106,55%	90,03%	85,67%
2	78,99%	80,81%	64,37%	77,67%	91,63%
3	83,81%	80,19%	112,85%	115,35%	111,28%
4	79,59%	75,96%	85,39%	85,26%	75,44%
5	64,29%	74,94%	68,64%	84,43%	88,66%
6	68,98%	73,47%	67,74%	54,58%	69,62%
7	47,06%	70,50%	95,87%	76,37%	81,05%
8	49,20%	71,78%	-43,85%	15,15%	3,55%
9	72,45%	68,49%	86,10%	75,50%	85,44%
10	69,19%	68,88%	93,31%	98,32%	99,68%
11	71,75%	68,77%	62,37%	83,97%	68,01%
12	77,71%	65,89%	194,78%	108,48%	134,31%

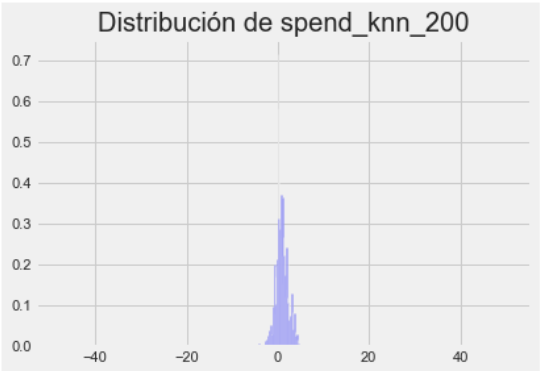
Tabla 9-10: *Uplift* en 'spend' en función de la variable 'Channel'

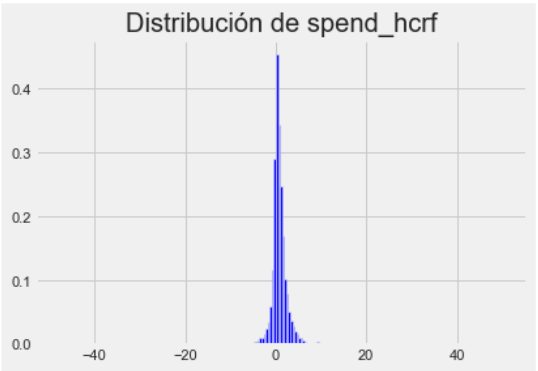
'Channel'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
Multichannel	94,01%	101,79%	104,09%	109,29%	112,70%
Phone	72,13%	70,92%	65,15%	57,40%	63,03%
Web	67,04%	70,97%	83,03%	96,77%	95,86%
Muestra de Testeo					
Multichannel	96,01%	102,35%	112,83%	116,79%	115,30%
Phone	73,41%	71,29%	67,06%	56,43%	62,09%
Web	68,58%	71,60%	91,63%	98,81%	98,87%

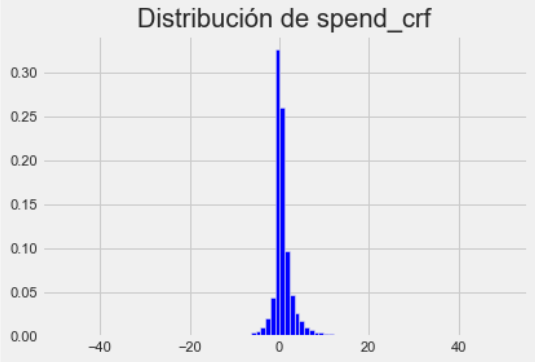
Tabla 9-11: *Uplift* en 'spend' en función de la variable 'newbie'

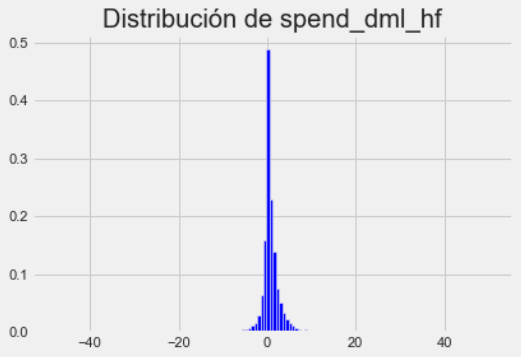
'newbie'	knn_275	ols	Crf	hcrf	dml_hf
Muestra de Entrenamiento					
0	71,13%	69,24%	62,32%	74,57%	75,08%
1	73,93%	80,11%	93,16%	87,62%	92,04%
Muestra de Testeo					
0	71,64%	69,19%	59,02%	74,77%	72,74%
1	76,41%	81,22%	107,90%	90,39%	96,98%

Figura 9-3: Distribución de las predicciones continuas del CATE en 'spend' en la población total (MEN_campaign)






Distribución de spend_crf
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−40
−20
0
20
40


Distribución de spend_dml_hf
0.5
0.4
0.3
0.2
0.1
0.0
−40
−20
0
20
40

10 Acrónimos

ATE: Average Treatment Effect

CATE: Conditional Treatment Effect

CRF: Causal Random Forest

DML: Double Machine Learning

HCRF: Honest Causal Random Forest

ITE: Individual Treatment Effect

KNN: K knearest kneigbours

MCO: Mínimos Cuadrados Ordinarios

OLS: Ordinary Least Squares

RCM: Rubin Causal Model

RF: Random Forest

SUTVA: Stable Unit Treatment Value Assumption

TOL: Transformed Outcome Loss

VI: Variable Importance

11 Referencias

[1] Athey, S. e Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. stat, 1050(5), 1-26.

[2] Athey, S. e Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27): 7353 – 7360.

[3] Baesens, B., Rösch, D. y Scheule, H. (2016). Credit Risk Analytics, Measurement Techniques, Applications, and Examples in SA. Wiley.

[4] Berrevoets, J., Verboven, S., y Verbeke, W. (2019). Optimising Individual-Treatment-Effect Using Bandits. In Neural Information Processing Systems (NeurIPS) 2019 workshop "Do the right thing": machine learning and causal inference for improved decision making Vancouver, Canada: Curran Associates, Inc.

[5] Breiman, L. (2001). Random Forests. Machine Learning. 45(1):5 32

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. y Newey, W. (2016). Double Machine Learning for Treatment and Causal Parameters. URL: *https://arxiv.org/abs/1608.00060*.

[7] Devriendt, F., Moldovan, D. y Verbeke, W., (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. In: *Big Data* 6 (2018), Nr. 1, S. 13–41. URL: https://doi.org/10.1089/big.2017.0104.

[8] Foster J. C., Taylor J. y Ruberg S. J. (2011). Subgroup identification from randomized clinical trial data. Statistics in Medicine, 30(24): 2867-2880

[9] Gross, R. T. (1993). Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988. Inter-university Consortium for Political and Social Research [distributor], 1993-10-03. URL: https://doi.org/10.3886/ICPSR09795.v1

[10] Hitsch, G. J., y Misra, S. (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. URL: https://ssrn.com/abstract=3111957

[11] Holland, P.W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945–960.

[12] LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4), 604-620. Retrieved April 26, 2021, from http://www.jstor.org/stable/1806062

[13] Lopez-Paz, D., y Oquab, M. (2016). Revisiting classifier two-sample tests. URL: https://arxiv.org/abs/1610.06545

[14] Louizos. C., Shalit, U., Mooij, J. M., Sontag, D. , Zemel, R. y Welling, M.. (2017). Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, páginas 6449–6459.

[15] Radcliffe, N. J. (2007), Using Control Group to Target on Predicted Lift: Building and Assessing Uplift Models. Direct Marketing Analytics Journal, 14-21. URL: http://docplayer.net/17470611-Using-control-groups-to-target-on-predicted-lift.html

[16] Radcliffe, N. J. y Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions.

[17] Rosenbaum P.R. y Rubin D.B. (1983). The central role of the propensity score in the observational studies for causal effects. Biometrika, 70(1): 41-55.

[18] Rubin D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688–701.

[19] Wager S. y Athey S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113:523, 1228-1242.

[20] Documentación CausalML – Python: https://causalml.readthedocs.io/en/latest/about.html

[21] Documentación de EconML – Python: https://econml.azurewebsites.net/index.html

[22] Documentación sklearn – Python: https://scikit-learn.org/stable/