Targeted Marketing

Isidoro Garcia

2021

Overview

Los contratan como data scientists para una empresa que vende electrodomesticos. La empresa lanzó un experimento de control aleatorio via un mail en donde se envió un catalogo de los productos al grupo de tratamiento `mailing_indicator`.

Tu objetivo es estimar el impacto del envío sobre el gasto incremental:

$$\tau_i = \mathbb{E}[Y_i(1) - Y_i(0)|\boldsymbol{x}_i],$$

En particular, queremos estimar el impacto de enviar el catalogo a nivel de cliente. Para ello, pondremos a competir algunos de los modelo de Causal Machine Learning que hemos aprendido en clase:

- Double Debiased Machine Learning
- Causal Forests

Adicionalmente, desarrollen una estrategia de focalización con base en los resultados de tu modelo. Elabora sobre la lógica económica (i.e. identifica los Beneficios y Costos Marginales de enviar la campaña). Finalmente, corrobora la validez externa de la estrategia usando datos de un año. Esto nos dará un termómetro de la utilidad del modelo para campañas posteriores.

Tip!: En los chunks donde vaya a haber modelos o cálculos complicados, usen `cache=T`

Paso 1: Estimación y predicción the Conditional Average Treatment Effects (CATE)

Carguemos los datos de 2015

```
library(tidyverse)
library(data.table)
library(gamlr)
library(grf)
library(xgboost)
library(ranger)
library(RCT)
library(lfe)
library(stargazer)
library(knitr)
```

```
load("Bases input/Customer-Development-2015.RData")
```

Dividimos la base en entrenamiento y validacion. Usamos un seed fijo para replicabilidad.

```
set.seed(1990)
crm<-
```

```
crm %>%
mutate(training_sample = rbinom(n = nrow(crm), 1, 0.7))
```

Data cleaning

1. Haz una primera revisión de la base. Cuantas variables tienen `NA`
2. Muestra la matriz de correlación entre variables. Muestra los pares de variables que tienen más de 95% de correlación. Remueve una de cada par multicolineal.

3 (2 pts). Corroba que la asignación tratamiento fue aleatoria mediante revisión del balance. Realiza las pruebas balance T y F. Cuántas variables salen desbalanceadas? Que muestra esto sobre la asignación de tratamiento?

4. Realize un ajuste de False Discovery Rate al 10%. Cuántas variables salen desbalanceadas ahora?

Estimación de impacto de tratamiento (ATE)

5 (2pts). Estima el impacto promedio de enviar el catalogo vía email. Estima el impacto sin controles y luego agregar dos estimaciones de robustez: 1) Agregando variables que salieron significativas y 2) Agregando variables que salieron significativas con el FDR. Interpreta los resultados

Estimación de efectos heterogeneos Usaremos el training sample para estimar el Conditional Average Treatment Effect de enviar el catalogo sobre el gasto en dólares. Estimaremos dos tipos de modelos (si agregan otro es bienvenido):

(a) Double Debiased LASSO
(b) Causal Forests

Separa la base de entrenamiento de la de validación

####Double Debiased LASSO

6 (3pts). Estima un Double Debiased LASSO. Asegurate de mostrar el código. (Tip: recuerda que necesitas guardar el LASSO de cada K para poder usarlo en la base de validación)

7 (2pts). Cuál es el impacto de tratamiento promedio? Estimalo de dos maneras: 1) `spend_resid~treat_hat + treat` y 2) `spend~treat_resid`. Sale lo mismo? Justifica tu respuesta

8 (3pts). Cuáles son las variables más importantes para las nuisance functions $T_i = g(X_i)+v_i$ y $y_i = m(X_i)+\epsilon_i$? (Tip: toma las variables que tengan $\beta \neq 0$ en cada k y haz un `inner_join`. De ahí muestra el promedio de los coeficientes) Interpreta la función $g(X_i)$, porque sale así?

9 (3pts). Ahora corre un DDML LASSO para encontrar los efectos a nivel cliente (Tip: interactúa todas las variables con `treat_resid`. Muestra el código. Qué varaibles salen significativas?

10 (2 pts). Predice el CATE en la base de entrenamiento y en la base de validación. Como se ve la distribución del impacto de tratamiento en ambas?

####Causal Forest

11 (2pts). Ahora vayamos al causal forest. Estima un causal forest en la base de entrenamiento (Estima 750 árboles)

12 (3pts). Cómo se distribuye el impacto de tratamiento? Cuál es el impacto de tratamiento (ATE)? Qué tanto se acerca al impacto de tratamiento "real"? Cómo se compara con el impacto estimado con el ddml simple?

13. Haz un scatter plot de las predicciones de ambos modelos? Hay alguna relación?

14 (4pts). Evalúa el poder predictivo de cada modelo (OOS). Esto se hace por modelo: Divide la muestra en 10 partes con base en el score de ddml. Para cada parte, estima el impacto de tratamiento vía una regresión y saca el promedio del score. Valida si para los grupos que dice el score el impacto será más grande, el coeficiente de la regresión es. Cómo se ven los modelos? Cuál parece ser mejor?

15 (6 pts). Construye una estrategia de focalización a nivel usuario con base a los resultados de cada modelo. Considera lo siguiente:

- El costo marginal de mandar el mail es 0.99 USD
- El Beneficio marginal es el impacto incremental la utilidad generada por esas ventas
- El margen de ganancia sobre las ventas es de 32.5 fijo

Con esto, indica:

- Cuantos usuarios entrarían a la campaña?
- A partir de cuánto lift (ventas incrementales) entran?
- Cuál es el impacto promedio esperado de tu población final?
- Cuánta utilidad haremos con esta estrategia? Cómo se compara con la utilidad de la campaña sin focalizar?

16 (3pts). Haz una gráfica del la utilidad total vs q (personas que entran en la campaña) para DDML y CF