Proyecto final

Machine Learning: Análisis contrafactual,
con Carlos Tabares

Technology & Data

Descripción general del proyecto:

El equipo de producto de tu empresa busca implementar una nueva campaña promocional con el objetivo de incrementar el volumen de ventas mediante la activación clientes nuevos que aún no han realizado su primera compra. Algunos miembros del equipo argumentan que una promoción tipo cashback sería más efectiva que otorgar un descuento. Además, consideran que esta campaña debería implementarse de manera recurrente sobre todos los clientes nuevos. Aún no estás seguro de que ésta sea la decisión correcta y que la promoción deba aplicar a todos los clientes, así que les propones llevar a cabo un experimento para probar ambas estrategias a fin de tomar una decisión correcta y focalizar los esfuerzos sobre los clientes de más alto valor.

Objetivo del proyecto:

Resolver un caso de negocio desde una perspectiva de Causal Data Science. Aprender cómo diseñar y medir una estrategia experimental y utilizar los resultados para desplegar campañas focalizadas que incrementen la rentabilidad del negocio.

Conocimientos a utilizar:

Análisis contrafactual, experimentación y causal machine learning.

Herramientas a utilizar:

El proyecto se desarrollará en R Studio y se entregará un R Markdown como documento final.

Descripción de avances:

Avance 1: Diseño Experimental y Asignación Aleatoria

Diseñarás un experimento desde su inicio mediante una asignación aleatoria estratificada. El equipo de comunicación te comparte una lista de los clientes interesados en tu producto que aún no han realizado una compra o se encuentran inactivos. Deberás segmentar a esta población en 3 grupos dependiendo del tipo de promoción que recibirán. Después de asignar tendrás que regresar la misma base de clientes agregando una columna con el identificador del grupo al que pertenece cada uno. Esto a fin de que puedan enviar las promociones correspondientes mediante su plataforma de distribución de correo.

Paso 1: Explora la base de datos de clientes nuevos. ¿A qué nivel de desagregación está la base? ¿Cuántos clientes únicos? ¿Qué variables tienen valores vacíos? Decide si debes excluir a esas observaciones o mantenerlas y justifica tu decisión.

Paso 2: ¿Qué variables crees que pudieran estar más correlacionadas con el impacto del tratamiento y generar sesgo? Recuerda considerar estas variables al momento de estratificar.

Paso 3: Realiza una asignación aleatoria de la población de clientes en 3 grupos de tamaño similar. El grupo asignado determinará el tipo de promoción que recibirán: 1) un grupo de control que no recibirá comunicación, 2) el grupo de tratamiento 1 que recibirá un correo con la promoción de un cash back y 3) el grupo de tratamiento 2 que recibirá el correo con la promoción de un descuento ¿Qué ventajas tiene hacer este diseño experimental en comparación con un AB test?

Paso 4: Realiza las pruebas de balance sobre todas las variables. ¿Están balanceadas las variables entre los 3 grupos?

Avance 2: Evaluación de impacto

Dos semanas después del envío de las promociones a la empresa le gustaría entender cuál fue el resultado del experimento, el impacto sobre el número de conversiones y el valor de las ventas promedio.

Deberás realizar una evaluación integral de la estrategia y determinar qué opción fue la más efectiva para incrementar las ventas y para qué perfil de clientes.

Paso 1: Realiza una comparación del porcentaje de conversiones y el valor promedio de las ventas entre los 3 distintos grupos de clientes. ¿Observas alguna diferencia entre los grupos?

Paso 2: Estima una regresión de evaluación de impacto de los efectos de tratamiento (ITT). Incluye efectos fijos por estrato en tu especificación. Reporta en una tabla el efecto promedio para cada grupo de tratamiento y su significancia estadística correspondiente.

Paso 3: Realiza la estimación de efectos heterogéneos usando las variables por las cuales estratificaste. ¿Observas alguna subpoblación por grupo de tratamiento para la cual los efectos difieran del promedio?

Paso 4: ¿Qué puedes concluir de la evaluación experimental? ¿Cuál sería tu recomendación para la empresa? ¿Vale la pena centrarse en un grupo específico de clientes?

Avance 3: Focalización de la estrategia

Finalmente, en el avance 3 utilizaremos un modelo de Causal Machine Learning para estimar el impacto de la promoción a nivel cliente, y responderemos a la pregunta de cuánto es el valor de una estrategia focalizada.

Revisando el presupuesto de la estrategia te das cuenta que asignar descuentos resultó costoso debido a que el valor de la ventas promedio y el porcentaje de recompra fueron más bajos de lo que se tenía previsto. No obstante, recuerdas que la evaluación del experimento mostró que algunos clientes reaccionaron de manera muy positiva a la promoción y puede existir una oportunidad de rentabilizar la estrategia si te centras únicamente en estos clientes. En este sentido, utilizarás los resultados del experimento para desarrollar un proyecto de focalización empleando modelos de inteligencia artificial causal.

Paso 1: Explora la base y asegúrate de tener todas tus variables en formato numérico. Si tienes variables de texto o factores, transfórmalas a variables categóricas. Si tienes variables con valores vacíos decide si debes excluir a esas observaciones o mantenerlas y justifica tu decisión.

Paso 2: Estima una matriz de correlaciones de todas tus variables. Muestra los pares de variables que tienen más de 95% de correlación y elimina una de cada par multicolineal.

Paso 3: Quédate únicamente con tus clientes del grupo de control y el tratamiento con el descuento. Divide aleatoriamente a la población en 2 muestras: la muestra de entrenamiento (70% de las observaciones) y la muestra de validación (30% de las observaciones)

Paso 4: Estima un causal forest en la base de entrenamiento (Estima 1000 árboles)

Paso 5: Realiza un histograma para mostrar la distribución del impacto de tratamiento? ¿Cuál es el impacto promedio de tratamiento estimado por el modelo? ¿Se asimila al efecto encontrado en el experimento?

Paso 6: Evalúa el poder predictivo del modelo en la base de validación. Recuerda dividir tu base de validación en k partes (k=10) con base en el score de predicción modelo. Posteriormente, estima el impacto de tratamiento (del experimento) en cada grupo de score y también calcula el promedio de las predicciones en cada grupo. (Tip: Puedes estimar el impacto de tratamiento con la función impact_eval considerando efectos heterogéneos por grupo de score). Valida si para los distintos grupos de score, la predicción del impacto promedio y el coeficiente de la regresión son crecientes y consistentes.

Paso 7: Predice cuál hubiera sido el impacto sobre las ventas si los clientes que recibieron el cashback hubieran recibido un descuento. Asume que estos clientes nunca fueron tratados y utiliza esta base para simular una estrategia de focalización a nivel usuario con base en los resultados de tu modelo. Considera que el monto de compra mínimo es de $7 y que sólo tienes presupuesto para dar 1,000 cupones de descuento. ¿Cuál es el impacto promedio y el impacto total esperado de los usuarios de tu campaña focalizada?

crehana