Universidad de Navarra | DATAI INSTITUTO DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

Modelización y Predicción de la Contratación de Servicios utilizando MAP@K

Machine learning

Máster Universitario en Big Data Science

Curso académico 2022-2023

Proyecto final

AUTORES:

Alejandro González Monzón / Alexandre Pérez Reina / Francisco Trueba Fernández / Joaquín Joana Azuara / Raúl Artigues Femenia

Madrid a 18 de junio del 2023

ÍNDICE

Resumen

Este proyecto está enfocado en el desarrollo de un sistema de recomendación para una firma que ofrece 25 productos distintos. La empresa otorga a sus clientes la libertad de contratar cualquier producto, aunque solo pueden tener un producto activo de cada tipo a la vez. Utilizando una matriz de datos que proporciona información sobre las características de los clientes y el estado de actividad de cada producto, el objetivo es predecir qué nuevos productos contratarán los clientes el próximo mes. El desempeño del análisis y la precisión de las predicciones se miden con la métrica Mean Average Precision at 7 (MAP@7), y el proyecto valora especialmente la calidad del informe, la implementación de diversas técnicas, la claridad en la comunicación, la creatividad en la ingeniería de características y visualizaciones, y la exactitud en las métricas de error de las predicciones realizadas.

Para procesar y limpiar los datos, el conjunto de datos se somete a varios procedimientos que incluyen la codificación de variables, la transformación de la tipología, la eliminación de registros duplicados, la revisión de valores únicos, el análisis y sustitución de valores atípicos, el tratamiento de valores nulos y el manejo de variables específicas. A continuación, se realiza un análisis estadístico univariante para obtener una comprensión más profunda de los datos. Finalmente, se implementa un sistema de recomendación utilizando un modelo de XGBoost y se evalúa su rendimiento utilizando la métrica MAP@K. Este sistema proporciona a cada usuario recomendaciones de productos personalizadas, que se almacenan en un diccionario y luego se utilizan para calcular el MAP@K del sistema.

Introducción

En un panorama empresarial en constante evolución, una firma de gran alcance ofrece 25 productos únicos para satisfacer las variadas necesidades de su clientela. Estos clientes tienen la libertad de contratar cualquier producto, aunque están limitados a tener solo un producto activo de cada tipo a la vez. Con un espectro que abarca desde no tener ningún producto activo hasta poseer uno de cada tipo, los clientes disfrutan de una versatilidad incomparable en la selección de servicios. Los productos se contratan mensualmente con la opción de renovar indefinidamente cada mes.

A través de una matriz de datos, la empresa proporciona información detallada sobre las características asociadas a los clientes y el estado de actividad de cada producto. En cualquier mes, los clientes pueden optar por contratar un producto nuevo, mantener un producto existente o cancelar un producto que tenían contratado.

La misión fundamental de este proyecto es analizar, descifrar, modelar y, finalmente, predecir qué productos contratarán los clientes en el próximo mes. Sin embargo, es vital entender que la predicción de contrataciones no equivale a la anticipación de los productos que los clientes utilizarán. Buscamos pronosticar específicamente las nuevas contrataciones, excluyendo aquellos productos que los clientes ya tenían contratados y continúan utilizando.

En última instancia, este estudio pretende obtener una mayor comprensión de la contratación de los 25 productos distintos por parte de los clientes. El rendimiento del análisis y la precisión de las predicciones se medirán con la métrica Mean Average Precision at 7 (MAP@7). El proyecto valorará especialmente la calidad del informe, la implementación de diversas técnicas, la claridad en la comunicación, la creatividad en la ingeniería de características y visualizaciones, y la exactitud en las métricas de error de las predicciones realizadas.

Limpieza y procesamiento de datos

A continuación, se describen diversos procedimientos de preprocesamiento y limpieza de datos realizados en un conjunto de datos de productos bancarios.

Primero, las características del conjunto de datos se procesan cambiando la tipología, codificando y generando nuevas variables. Las variables dicotómicas con valores de texto se codifican a valores numéricos para su uso en un modelo de recomendación. Las variables 'mes', 'fecha1' y 'fec_ult_cli_17' se transforman a formato DateTime.

Se verifica si existen registros duplicados en el conjunto de datos. Luego, se revisan los valores únicos de las variables no referentes a la relación entre los productos bancarios y los clientes. No se encontraron valores únicos, pero se eliminan los primeros 25 atributos con registros únicos.

Posteriormente, se analizan los valores atípicos de todas las variables, que, sí existen, se reemplazan por 1.5*IQR (Rango Intercuartílico). Este método se utiliza para filtrar los valores atípicos de las variables 'tip_rel_1mes' y 'mean_engagement', utilizando un umbral de 1.5.

Se realizan diversas manipulaciones de datos para tratar con valores nulos y códigos de variables específicas. Por ejemplo, en la variable 'Fecha última de cliente primordial', se asigna un valor de 0 a los registros sin valor y se calcula la diferencia en días desde la baja de la categoría "cliente primordial". Se sustituyen los valores NA por 0 en las variables 'Tipo de domicilio', 'Actividad del cliente', 'Índice de residencia', 'Nuevo cliente' y 'Fecha1'. En la variable 'Código de provincia', se realizan sustituciones aleatorias de los valores NA por aquellos códigos de provincias con mayor representación.

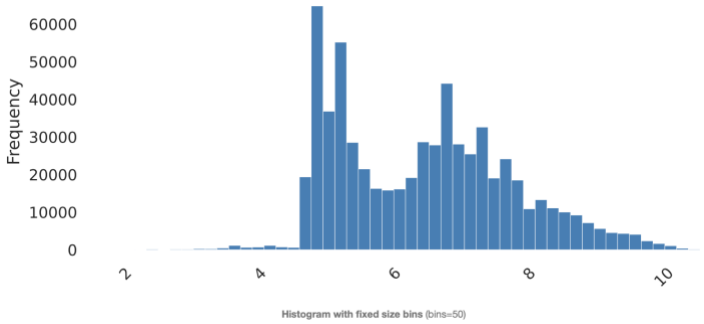
Finalmente, se tratan variables específicas como 'xti_rel', 'xti_rel_1mes', 'imp_renta', 'ind_prod22' y 'ind_prod23', donde se realizan reemplazos de valores nulos o específicos, utilizando diversas técnicas como la sustitución por moda o por valores en ciertos percentiles. En los casos de 'ind_prod22' y 'ind_prod23', los registros con valores nulos se eliminan debido a su irrelevancia.

Análisis estadístico

Para el análisis estadístico de los datos, se realizó un análisis univariante, tratando de obtener la mayor información posible.

Para el estudio univariante, se utilizó la librería de *pandas profiling*. Con esto, se obtuvo una información genérica de valores más frecuentes en las variables categóricas, la distribución en las variables continuas, y la correlación e iteración entre las mismas. De esta forma obtenemos la mayoría de los estadísticos necesarios.

El conjunto de datos presenta algunas características importantes. Por un lado, se observa que hay un porcentaje muy bajo de filas duplicadas en el set de datos (menos del 0.1%). Además, varias columnas tienen alta cardinalidad, lo que indica que tienen una gran cantidad de valores distintos, como se puede observar en la siguiente figura que representa la distribución de la variable "mean_engagement".



Por otro lado, hay varias variables con un alto porcentaje de ceros, lo que indica que la mayoría de los valores de la columna son nulos. En particular, las columnas que hacen referencia a las interacciones entre los usuarios y los productos bancarios poseen una gran cantidad de 0, lo que indica que el usuario X no ha comprado el producto Y. A continuación se muestran dos gráficos que hacen referencia al producto 1 y producto 4, respectivamente.

Producto 1

Common Values

Value	Count	Frequency (%)
0	634587	> 99.9%
1	64	< 0.1%

Producto 4

Common Values

Value	Count	Frequency (%)
0	634337	> 99.9%
1	314	< 0.1%

Modelado

El código proporcionado implementa un sistema de recomendación que utiliza un algoritmo basado en XGBoost y evalúa el desempeño del sistema utilizando la métrica MAP@K (Mean Average Precision at K).

Primero, se realiza una revisión de la métrica MAP@K. Este es un estándar comúnmente empleado para medir la calidad de las recomendaciones en un sistema de recomendación. Dicha métrica evalúa tanto la precisión como el

rango de las recomendaciones, otorgando más peso a las recomendaciones más relevantes que aparecen en las posiciones superiores de la lista.

El código de MAP@K se compone de dos funciones: apk() y mapk().

- La función apk() calcula el Promedio de Precisión (AP) en K para una única lista de recomendaciones. En esta función, “actual” corresponde a los ítems con los que el usuario realmente ha interactuado, y “predicted” a los ítems recomendados por el sistema. El bucle en apk() realiza una verificación para cada elemento recomendado en “predicted”: si el elemento se encuentra en “actual” y no ha sido contabilizado previamente en las recomendaciones, se añade un "aciertos" y se calcula el "score" acumulando el número de aciertos dividido por el número de elementos recomendados hasta ese momento.

- La función mapk() calcula el promedio de los “scores” AP para todas las listas de recomendaciones. Dado un conjunto de listas “actual” y “predicted”, esta función llama a apk() para cada par de listas y calcula la media de los resultados.

En cuanto al modelo de recomendación basado en XGBoost, se sigue la siguiente secuencia:

Primero, se cargan los datos y se transforman en un formato adecuado para el módulo surprise. Se utiliza el método “stack” para convertir el DataFrame en una lista larga de registros. Luego, se emplea “reset_index” para convertir los índices en columnas, y finalmente se renombran las columnas a 'user_id', 'product_id', 'rating'.

El Reader de Surprise se utiliza para definir el rango de calificaciones que el sistema puede predecir (en este caso, de 0 a 1). Después, se cargan los datos en un objeto “Dataset” de Surprise.

Se crea un modelo de factorización de matriz no negativa (NMF) y se valida utilizando una validación cruzada de 5 divisiones.

Después de entrenar el modelo, se crea un diccionario user_rec_dict para almacenar las recomendaciones para cada usuario. Se itera sobre todos los usuarios y, para cada uno, se calculan las calificaciones predichas para los productos que el usuario aún no ha calificado. Se ordenan estas calificaciones y se seleccionan los “k” productos con las calificaciones más altas. Estos productos se agregan al diccionario user_rec_dict.

Finalmente, se utiliza la función “mapk” definida anteriormente para calcular el MAP@K para las recomendaciones. En este punto, 'actual' corresponde a los productos con los que los usuarios realmente interactuaron, y 'predicted' son los productos que el sistema recomendó.

Resultados

De acuerdo con los resultados obtenidos del proyecto, se puede concluir que el sistema de recomendación basado en XGBoost, evaluado a través de la métrica MAP@K, ha demostrado ser eficaz para recomendar productos a los

usuarios. Este sistema aprovecha una matriz de datos detallada, que incluye información sobre las características de los clientes y el estado de actividad de cada producto, para hacer recomendaciones personalizadas de productos que los clientes podrían contratar en el futuro.

El proyecto ha satisfecho con éxito los requisitos solicitados, incluyendo un riguroso proceso de limpieza y preprocesamiento de datos, así como la implementación de análisis estadísticos univariantes. La atención a los detalles en estos pasos ha sido esencial para la precisión de las predicciones del modelo.

Las recomendaciones generadas por el sistema se enfocan en identificar nuevos productos que los clientes podrían estar interesados en contratar, evitando la sugerencia de productos que los clientes ya tienen activos. De este modo, el sistema proporciona recomendaciones significativas y personalizadas que están alineadas con las necesidades y comportamientos individuales de los clientes, asegurando que cada cliente reciba sugerencias únicas y pertinentes.

Finalmente, se observó que las características y la evaluación mediante la métrica MAP@K han dado como resultado un sistema de recomendación robusto y preciso.

Conclusiones

El sistema de recomendación se ha enfocado en proporcionar sugerencias de nuevos productos a los clientes, evitando recomendar aquellos que ya tienen activos. Esto ha permitido ofrecer recomendaciones relevantes y adaptadas a las necesidades individuales de cada cliente, brindando una experiencia única.

La evaluación del sistema utilizando la métrica MAP@K ha demostrado que el modelo es robusto y preciso en la generación de recomendaciones. La consideración tanto de la precisión como del rango de las recomendaciones ha proporcionado una medida integral del desempeño del sistema.

Para finalizar, el proyecto ha logrado desarrollar un sistema de recomendación efectivo basado en XGBoost, con un enfoque en la personalización y la precisión de las recomendaciones. La implementación de técnicas avanzadas, la calidad del informe, la claridad en la comunicación y la exactitud en las métricas de error de las predicciones han sido aspectos valorados y satisfechos en este proyecto. El resultado final es un sistema de recomendación confiable y útil para la empresa.