PROPUESTA PARA MONOGRAFÍA

HADYS OSVALDO AGUDELO

CC. 71721484

JUAN PRADO ESCOBAR

CC. 1017209468

ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

2023

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

TÍTULO DEL PROYECTO: Predicting Credit Card Customer Segmentation

INTEGRANTES (NOMBRES COMPLETOS Y EL REPOSITORIO DE GITHUB DE CADA UNO): Hadys Osvaldo Agudelo -- https://github.com/osvalcode
Juan Prado Escobar – https://github.com/juanprado19

ASESOR: Efrain Oviedo

DESCRIPCIÓN DEL PROBLEMA PREDICTIVO A RESOLVER

Segmentación de clientes basado en datos recopilados a partir de tarjetas de crédito.

- URL de la base de datos seleccionada

https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m

- Breve descripción del dataset

El dataset contiene amplia información de cartera de clientes de tarjeta de crédito, con el objetivo de predecir la pérdida de usuarios por parte una empresa. Este, incluye detalles tal como edad, sexo, estado civil y categoría de ingresos, así como información sobre la relación de cada cliente con el proveedor de la tarjeta de crédito, número de meses que han transcurrido desde la transacción (préstamo) y los períodos de inactividad, además de datos del comportamiento de gastos de los diferentes usuarios. A partir de este tipo de datos, se podría capturar información de interés que ayude a determinar la estabilidad de la cuenta a largo plazo y así a futuro poder brindar soluciones a clientes individuales según el comportamiento que se analice basado en los propios datos del usuario.

MÉTRICAS DE MACHINE LEARNING (DESEMPEÑO)

La elección de métricas de desempeño dependerá del comportamiento y las propias características de los datos. Por lo que se podrían utilizar algunas de las siguientes:

- **Silhouette Score:** esta mide la similaridad de cada punto respecto a su propio clúster en comparación con otros clústeres. Un valor cercano a 1 indica un buen *clustering* (agrupamiento), mientras que uno cerca a -1 indica un bajo *clustering*.

- **Calinski-Harabasz Index:** medida de la relación entre la varianza entre clústeres y la varianza dentro del clúster. Una puntuación alta indica mejor agrupación.

- **Davies-Bouldin Index:** medida de similitud prmedio entre cada clúster y su clúster más similar. Una puntuación más baja indica una mejor agrupación

- **Puntuaciones homogeneidad, integridad y medida -V (V-Measure):** estas métricas se utilizan para medir la calidad de los resultados del agrupamiento frente a etiquetas de conocidas ground truth. Dado que este es un problema de aprendizaje no supervisado, es posible que no se tenga acceso a estas etiquetas, pero aún así se pueden usar para comparar diferentes algoritmos de agrupación

- **Rand Index:** mide la similaridad entre los true *labels* y los predichos (en el caso de que se tenga acceso a los ground truth labels). Su puntuación va de 0 a 1, donde 1 indica una agrupación perfecta.

Asimismo, existen otras métricas, tales como Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index (FMI) y Jaccard Index, las cuales se pueden utilizar si se conocen los *true-labels* del problema en cuestión.

MÉTRICA DE NEGOCIO

Para este tipo de problema en el que se implementarán técnicas de aprendizaje no supervisado, específicamente para la segmentación de clientes basado en el comportamiento de los usuarios con las tarjetas de crédito, se podría utilizar métricas de negocio (las cuales están relacionas con los objetivos de las organizaciones y el impacto general del modelo en el negocio específico) como el incremento de la fidelidad del cliente con la tarjeta, mejora en la tasa de retención de clientes, incremento en la satisfacción del usuario, reducción en la tasa de abandono o pérdida de clientes, aumento de los ingresos por ventas, mejora en la eficiencia operacional, etc.

Es importante resaltar que la elección de la métrica de negocio depende de la usabilidad que se le dé al modelo en la organización y el objetivo que se quiera medir en específico. Estas métricas deben ser relevantes y medibles y a su vez utilizadas para evaluar el impacto general del modelo en el negocio.

CRITERIO DE DESEMPEÑO DESEABLE EN PRODUCCIÓN

El desempeño deseable en producción depende del uso que se le vaya a dar al modelo y de los objetivos que se pretendan medir a partir de la segmentación obtenida. Por ejemplo, si se utilizara el modelo entrenado (el cual debería estar con un buen rendimiento, brindando confiabilidad en que ese cliente pertenece a ese tipo de clúster en específico) para medir si hubo una mejora en la tasa de retención de cliente, el criterio de desempeño deseable sería que según la estrategia que se haya elegido para aplicar a los clientes, esta esté mejorando con el tiempo, indicando que los usuarios que utilizan dichas tarjetas de crédito hayan incrementado su fidelidad con la entidad bancaría, disminuyendo la tasa de deserción. Para esto, también es de importancia entrenar el algoritmo constantemente y revisar de forma periódica el desempeño de los objetivos de la organización planteados.