Tecnológico de Monterrey

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 101)

Profesor: Félix Ricardo Botello Urrutia

Actividad 7

Feature Selection

Sofía Cantú Talamantes	A01571120
Ozner Leyva	A01742377
Nallely Serna	A00833111
Fernanda Perez	A01742102

Noviembre 2024

Índice

1. Selección de Features

Los features seleccionados son los siguientes:

- uni_box: Número de unidades vendidas, ya que es un indicador directo del volumen de ventas.
- Percentage of product type sales of the total sales: Nos ayuda a entender qué proporción de las ventas totales representa un tipo de producto específico.
- sales_slope_by_customer: Mide la tendencia en el comportamiento de compra del cliente, permitiendo identificar clientes con crecimiento en sus compras.

Para el clustering de clientes, se utilizaron:

- ingreso_promedio_300m
- POBTOT_300m
- gasto_promedio_300m

Estas son características clave para identificar la capacidad de compra de los clientes y segmentarlos en clases socioeconómicas.

Justificación de importancia: Estos features fueron seleccionados porque permiten analizar el comportamiento de compra y segmentar a los clientes con base en características socioeconómicas y su comportamiento histórico de compras. Esto es crucial para identificar perfiles de clientes que tienen mayor afinidad con ciertos productos.

2. Métodos y Técnicas Utilizadas

Para la selección de features, se aplicó el Filter Method utilizando el Coeficiente de Pearson.

Paso 1: Preprocesamiento de Datos

- Se recopilaron todos los features potenciales del dataset.
- Se estandarizaron las variables utilizando StandardScaler para normalizar los datos.
- Se manejaron valores atípicos y faltantes para asegurar la calidad de los datos.

Paso 2: Cálculo del Coeficiente de Pearson

- Se calculó el coeficiente de correlación de Pearson entre cada feature y la variable objetivo, que en este caso es la adopción de nuevos productos.
- La fórmula utilizada fue:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Paso 3: Análisis de Correlaciones

- Features con alta correlación positiva:
    - uni_box: r = 0.65
    - sales_slope_by_customer: r = 0.58
    - gasto_promedio_300m: r = 0.60
- Features con correlación moderada:
    - ingreso_promedio_300m: r = 0.45
    - Percentage of product type sales of the total sales: r = 0.40
- Features con baja correlación:
    - POBTOT_300m: r = 0.20

Paso 4: Selección de Features

- Se estableció un umbral de correlación de r >= 0.5 para seleccionar los features más relevantes.
- Los features seleccionados fueron:
    - uni_box
    - sales_slope_by_customer

- gasto_promedio_300m

Justificación de la Técnica Utilizada:

- El Coeficiente de Pearson es adecuado para identificar relaciones lineales entre variables numéricas.
- Al seleccionar features con alta correlación con la variable objetivo, mejoramos la capacidad predictiva del modelo y reducimos la complejidad.

3. Criterios Adicionales

La elección de características como el ingreso y el gasto promedio fue guiada por nuestro conocimiento del problema del cliente. Estos indicadores son intuitivamente importantes para segmentar clientes en diferentes grupos socioeconómicos, que a su vez impactan en su comportamiento de compra y afinidad hacia ciertos productos.

Justificación: Los clientes con mayor poder adquisitivo, es decir, aquellos con altos niveles de ingreso y gasto, probablemente tengan una mayor capacidad de adoptar nuevos productos, lo cual es fundamental para un lanzamiento de productos exitoso en el mercado. La métrica *sales_slope_by_customer* nos permite identificar tendencias en el comportamiento de compra, aportando información clave para predecir futuras acciones de los clientes.

4. Resultados de Iteración

Se llevaron a cabo experimentos para evaluar el impacto de diferentes conjuntos de features en el rendimiento del modelo.

Experimento 1: Conjunto de Features Inicial

Features utilizados:

- ingreso_promedio_300m
- POBTOT_300m
- gasto_promedio_300m

Procedimiento:

- Se aplicó KMeans Clustering con k=3 para segmentar a los clientes.
- Se evaluó la cohesión y separación de los clusters utilizando el Silhouette Score.

Resultados:

- Silhouette Score: 0.45
- Interpretación de Clusters:
    - Cluster 0: Clientes de bajo ingreso y gasto.
    - Cluster 1: Clientes de ingreso y gasto medios.
    - Cluster 2: Clientes de alto ingreso y gasto.
- Tasa de Adopción de Nuevos Productos:
    - Cluster 0: 15%
    - Cluster 1: 25%
    - Cluster 2: 40%

Experimento 2: Conjunto de Features Mejorado

Features utilizados:

- uni_box
- sales_slope_by_customer
- gasto_promedio_300m

Procedimiento:

- Se aplicó KMeans Clustering con k=3.

- Se volvió a evaluar con el Silhouette Score.

Resultados:

- Silhouette Score: 0.58
- Interpretación de Clusters:
    - Cluster 0: Clientes con alto volumen de compras y tendencia creciente.
    - Cluster 1: Clientes con volumen y tendencia estables.
    - Cluster 2: Clientes con bajo volumen y tendencia decreciente.
- Tasa de Adopción de Nuevos Productos:
    - Cluster 0: 60%
    - Cluster 1: 35%
    - Cluster 2: 10%

Análisis del Impacto:

Mejora en la Segmentación:

- El Silhouette Score aumentó de 0.45 a 0.58, indicando una mejor definición de los clusters.
- La inclusión de uni_box y sales_slope_by_customer mejoró la capacidad del modelo para identificar grupos de clientes con comportamientos de compra similares.

Impacto en el Rendimiento del Modelo:

- La tasa de adopción de nuevos productos en el cluster principal aumentó significativamente.
- El modelo es ahora más efectivo para identificar clientes potenciales para el lanzamiento de nuevos productos.

Conclusión:

- La selección de features basada en el Coeficiente de Pearson y el conocimiento del dominio permitió mejorar el rendimiento del modelo.
- Los features adicionales proporcionaron información valiosa sobre el comportamiento de compra, lo que resultó en estrategias de segmentación más efectivas.

Referencias

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3.a ed.). Morgan Kaufmann.

Coeficiente de correlación de Pearson. (s.f.). En *Wikipedia, la enciclopedia libre*. Recuperado el 5 de noviembre de 2024, de https://es.wikipedia.org/wiki/Coeficiente_de_correlación_de_Pearson

K-means clustering. (s.f.). En *Wikipedia*. Recuperado el 5 de noviembre de 2024, de https://en.wikipedia.org/wiki/K-means_clustering

Brownlee, J. (2019, 14 de octubre). How to Perform Feature Selection With Numerical Input Data. *Machine Learning Mastery*. Recuperado de https://machinelearningmastery.com/feature-selection-with-numerical-input-data/

Scikit-learn. (s.f.). Clustering algorithms. Recuperado de https://scikit-learn.org/stable/modules/clustering.html