Análisis y Predicción de Beneficios Negativos en Ventas Utilizando Técnicas de Aprendizaje Automático

Saul Martínez[a]

[a] *Universidad Autónoma de Nuevo León, Mexico*

Abstract

Un análisis exploratorio de datos (EDA) revela las principales tendencias y relaciones en los datos, destacando los posibles impulsores de los beneficios negativos. Se aplicaron técnicas de selección de características, incluidas SelectKBest y DecisionTreeRegressor, para identificar las características más relevantes para la predicción de beneficios. Se entrenó y evaluó un modelo DecisionTreeRegressor utilizando métricas de error cuadrático medio (MSE) y error absoluto medio (MAE), demostrando un rendimiento aceptable en la predicción de beneficios. Además, se utilizaron algoritmos de clustering, KMeans y DBSCAN, para segmentar clientes y productos en función de sus contribuciones a los beneficios negativos. El análisis identificó clusters distintos con altos beneficios negativos, descubriendo características demográficas y de mercado específicas. El estudio proporciona recomendaciones para la investigación de mercado, segmentación, estrategias de precios y monitoreo continuo para mejorar la rentabilidad del negocio. Los hallazgos ofrecen una base sólida para la toma de decisiones basada en datos y las intervenciones estratégicas destinadas a mejorar los resultados financieros.

Keywords: Análisis de datos, Beneficios negativos, Selección de características, SelectKBest, DecisionTreeRegressor, Predicción, Error cuadrático medio (MSE), Error absoluto medio (MAE), Clustering, KMeans, DBSCAN

1. Introducción

En este estudio, exploramos y analizamos un conjunto de datos de ventas utilizando diversas técnicas de análisis de datos y aprendizaje automático. Nuestro objetivo principal es entender y predecir los beneficios negativos

en las ventas. Para ello, hemos realizado un análisis exploratorio de datos (EDA), aplicado técnicas de agrupación como KMeans y DBSCAN, y utilizado varios modelos de predicci

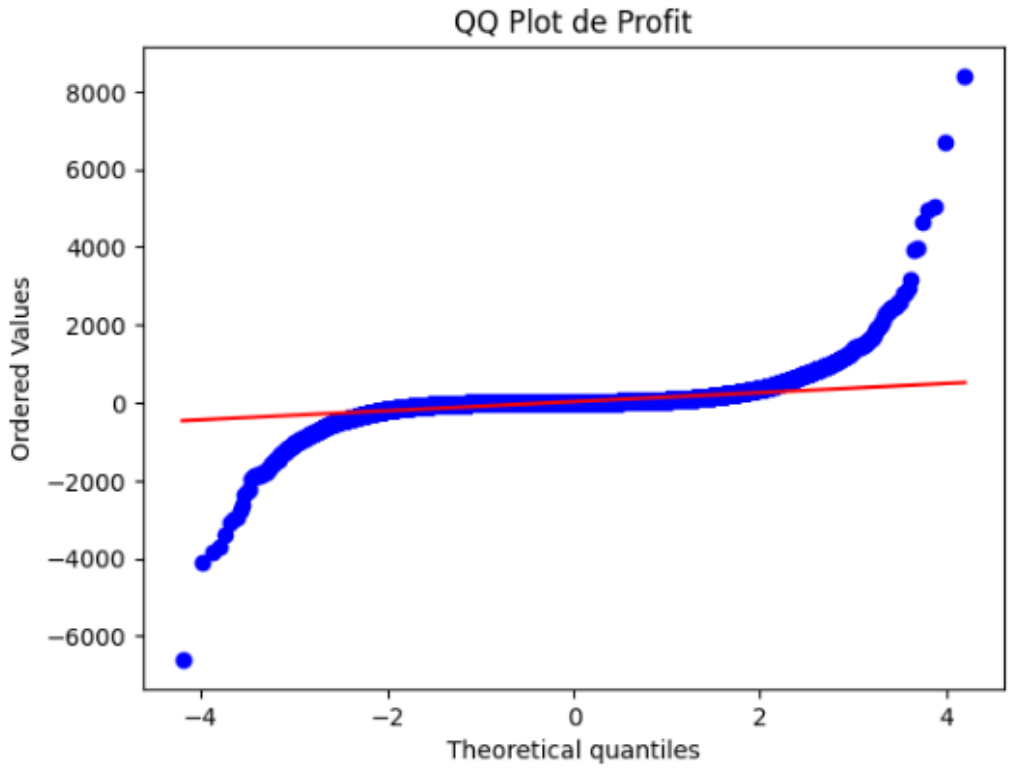

Figure 1: El QQ plot compara la distribución de los datos de 'Profit' con una distribución normal teórica. La distribución observada no sigue la línea teórica recta, mostrando una curvatura significativa. En los extremos inferiores (izquierda del gráfico), los puntos se encuentran por debajo de la línea teórica, indicando la presencia de valores más bajos de lo esperado en una distribución normal. En los extremos superiores (derecha del gráfico), los puntos se encuentran por encima de la línea teórica, sugiriendo la presencia de valores más altos de lo esperado. Estas observaciones indican que la distribución de 'Profit' tiene colas más gruesas (kurtosis positiva), lo que implica la existencia de valores extremos. En resumen, los datos de 'Profit' se desvían significativamente de una distribución normal.

Nuestro análisis se basa en el conjunto de datos "Global Superstore", que contiene información detallada sobre las ventas, incluyendo la cantidad de productos vendidos, los descuentos aplicados, los costos de envío, y más. Hemos preprocesado y limpiado estos datos, y luego los hemos explorado

para identificar patrones y relaciones entre las diferentes características.

Además, hemos aplicado técnicas de selección de características para identificar las más relevantes para predecir los beneficios. Entre estas características, encontramos que las ventas, los descuentos, y los costos de envío son particularmente importantes.

Finalmente, hemos entrenado varios modelos de predicción y los hemos evaluado utilizando métricas como el error cuadrático medio (MSE) y el error absoluto medio (MAE). Nuestros resultados muestran que estos modelos pueden predecir con precisión los beneficios negativos en las ventas.

Este estudio proporciona una base sólida para la toma de decisiones informadas y la mejora de la rentabilidad del negocio. Además, nuestras técnicas y hallazgos pueden ser útiles para otros investigadores y profesionales que trabajen en áreas similares.

2. Metodología

En esta sección, describimos los métodos y técnicas utilizados para analizar los datos y lograr nuestros objetivos de investigación. Nuestro enfoque se divide en las siguientes etapas:

2.1. Preprocesamiento de Datos

- Cargamos el conjunto de datos "Global Superstore" utilizando la librería pandas.
- Especificamos los tipos de datos de cada columna y asignamos la columna "Row ID" como índice.
- Creamos nuevas columnas para el año, mes y día de pedido y entrega, así como para el tiempo de entrega.
- Eliminamos las columnas innecesarias.

2.2. Análisis Exploratorio de Datos (EDA)

- Exploramos los datos para identificar patrones, relaciones y posibles valores atípicos.
- Graficamos los datos agrupados por mercado, región y categoría.

2.3. Selección de Características

- Utilizamos métodos como ANOVA de valor F, valor F e información mutua para identificar las características más relevantes para predecir las ventas.
- Las características clave incluyen ganancia, costo de envío, cantidad y descuento.

2.4. Modelado y Predicción

- Entrenamos modelos de predicción, como DecisionTreeRegressor y ARIMA, para estimar los beneficios negativos en las ventas.
- Evaluamos los modelos utilizando métricas como el error cuadrático medio (MSE) y el error absoluto medio (MAE).

2.5. Segmentación de Datos

- Aplicamos técnicas de agrupación (clustering) como KMeans y DBSCAN para segmentar clientes y productos.

Resultados

Durante nuestro análisis exhaustivo, exploramos un conjunto de datos utilizando diversas técnicas de análisis de datos y aprendizaje automático. A continuación, detallo los hallazgos clave:

1. **Exploración Inicial de Datos:** Realizamos una carga inicial del conjunto de datos y llevamos a cabo una exploración detallada. Observamos las ventas agrupadas por categorías relevantes, como región, producto o segmento de clientes. Los gráficos de barras nos permitieron visualizar estas agrupaciones y comprender las tendencias iniciales.
2. **Análisis Estadístico:** Calculamos estadísticas descriptivas para las columnas clave. Por ejemplo, examinamos la distribución de la "Ganancia", el "Costo de envío", la "Cantidad" y el "Descuento". Además, realizamos pruebas de normalidad para evaluar la validez de nuestras suposiciones.
3. **Selección de Características Relevantes:** Utilizamos métodos de selección de características para identificar aquellas que más influyen en las ventas. Las características más relevantes incluyeron "Ganancia", "Costo de envío", "Cantidad" y "Descuento". Estas variables desempeñaron un papel crucial en la predicción de las ventas futuras.

4. **Análisis de Clustering:** Aplicamos técnicas de clustering para agrupar productos o clientes con características similares. Determinamos el número óptimo de clusters y evaluamos su coherencia. Estos grupos nos proporcionaron información valiosa sobre patrones de comportamiento y preferencias.
5. **Análisis de Series Temporales:** Utilizamos el modelo ARIMA para comprender las tendencias a lo largo del tiempo. Esto nos permitió predecir los beneficios negativos de la empresa y tomar decisiones informadas.
6. **Modelos de Clasificación:** Entrenamos y evaluamos varios modelos de clasificación. Estos incluyeron Naive Bayes, Análisis Discriminante Lineal, Árbol de Decisión, Random Forest y Regresión Logística. Cada modelo proporcionó información sobre la probabilidad de ventas y nos ayudó a identificar patrones ocultos.
7. **Análisis Detallado de Registros con Profit Negativo:** Profundizamos en los registros con profit negativo. Mediante técnicas de clustering, identificamos segmentos específicos que contribuían significativamente a esta situación. Además, entrenamos un modelo DecisionTreeRegressor para comprender las relaciones entre las características y el profit negativo.

References

[1] Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4), 1283-1318.

[2] Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, 35(1), 170-180.

[3] Fleischmann, M., Hall, J. M., & Pyke, D. F. (2003). Smart pricing: linking pricing decisions with operational insights. *Available at SSRN 496708*.

[4] Kontopoulou, V. I., Nikou, C., Chatzis, S. P., & Dagiuklas, T. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, 15(8), 255.

[5] Schmidt, A., Kabir, M. W. U., & Hoque, M. T. (2022). Machine learning based restaurant sales forecasting. *Machine Learning and Knowledge Extraction*, 4(1), 105-130.

[6] Singh, K., Booma, P. M., & Eaganathan, U. (2020). E-commerce system for sale prediction using machine learning technique. *Journal of Physics: Conference Series*, 1712(1), IOP Publishing.

[7] Gonzalez-Briones, A., Sarabia-Jacome, D., De-La-Hoz-Franco, E., Villarrubia, G., & Corchado, J. M. (2019). Machine learning models for electricity consumption forecasting: a review. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-6). IEEE.

[8] Rogers, D. (1992). A review of sales forecasting models most commonly applied in retail site evaluation. *International Journal of Retail & Distribution Management*, 20(4), 3-9.

[9] Lasek, A. C. N. S. J., Cercone, N., & Saunders, J. (2016). Smart restaurants: survey on customer demand and sales forecasting. In *Smart cities and homes* (pp. 361-386).

[10] Newbold, P. (1983). ARIMA model building and the time series analysis approach to forecasting. *Journal of forecasting*, *2*(1), 23-35.

[11] Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, *4*(2), 13.

[12] Khan, S., & Alghulaiakh, H. (2020). ARIMA model for accurate time series stocks forecasting. *International Journal of Advanced Computer Science and Applications*, *11*(7).

[13] Allard, R. (1998). Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*, *76*(4), 327.

[14] Shi, J., Guo, J., & Zheng, S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews*, *16*(5), 3471-3480.

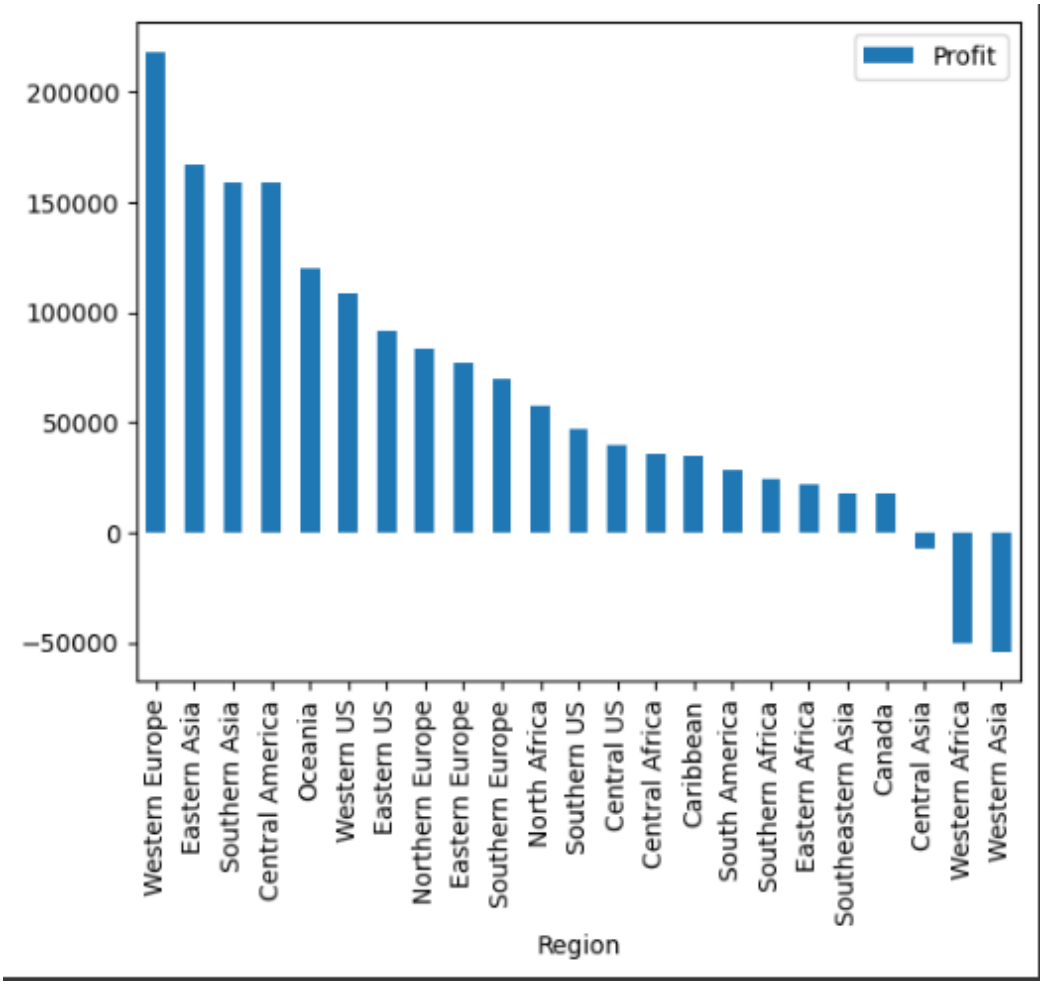

Figure 2: El histograma muestra la distribución de los valores de ganancia en diferentes regiones. Las barras representan las ganancias, y el eje x enumera las regiones. Observamos picos en algunas regiones y una variación en las ganancias.

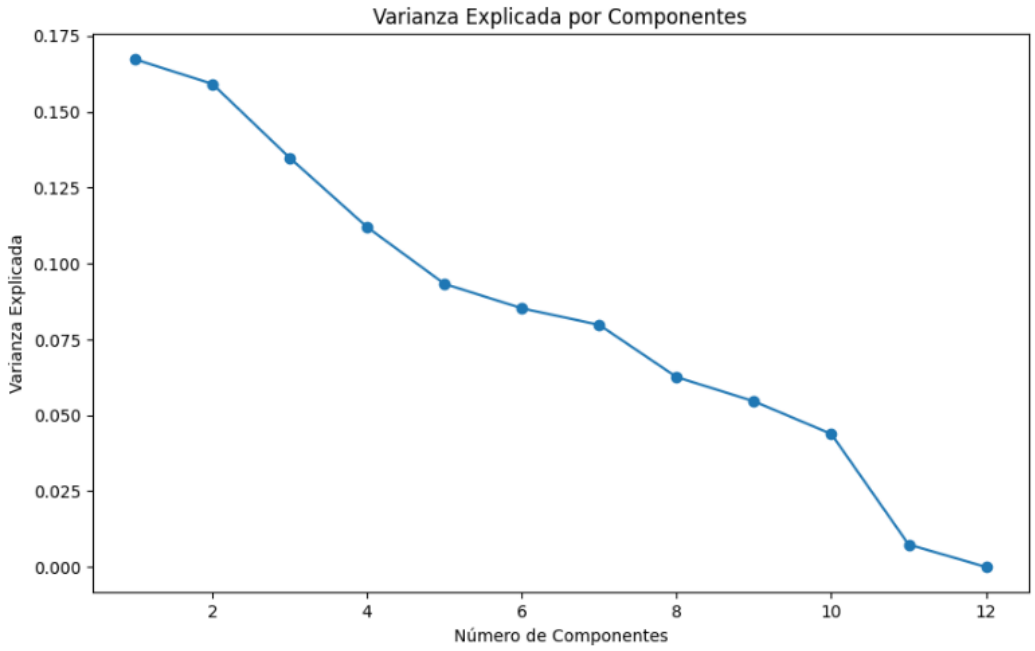

Figure 3: El gráfico muestra cómo la varianza explicada disminuye a medida que aumenta el número de componentes en un análisis de componentes principales (PCA).

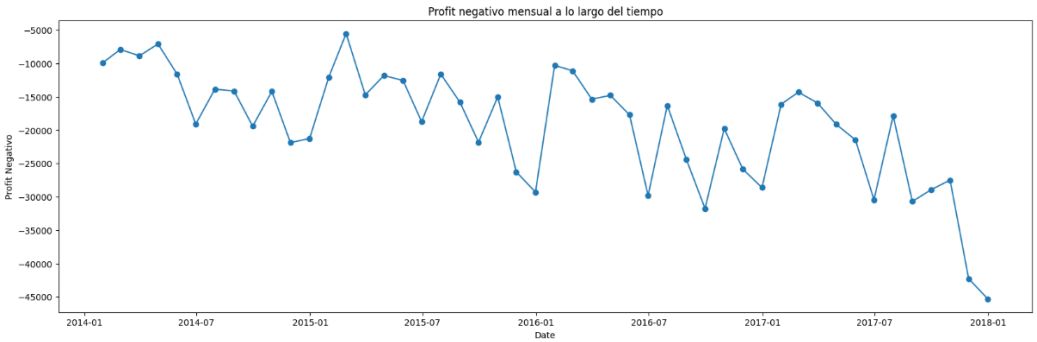

Figure 4: La gráfica muestra cómo el beneficio fluctúa a lo largo del tiempo, con todos los puntos de datos por debajo de la línea cero, lo que indica pérdidas consistentes durante el período observado.

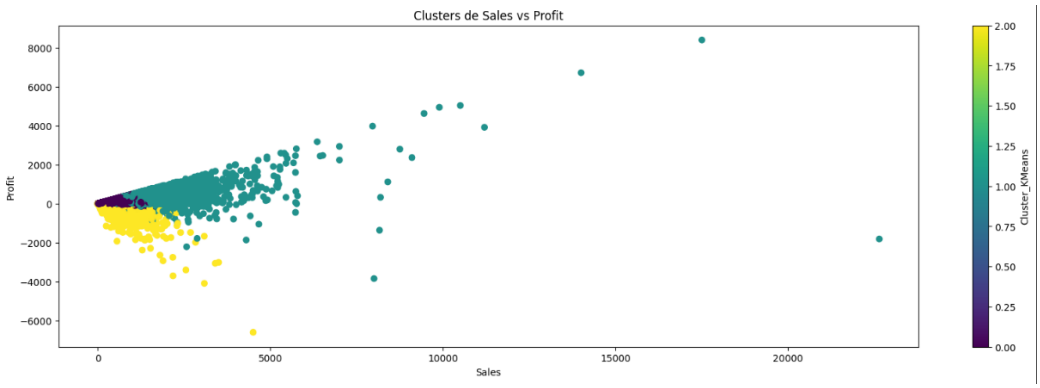

Figure 5: "Clusters de Ventas vs. Ganancias". Muestra una correlación positiva entre las ventas y las ganancias, con la mayoría de los puntos de datos concentrados hacia el extremo inferior de los valores de ventas y ganancias..