Tecnológico de Monterrey

Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Antoine Ganem Nuñez A01644024

14 de septiembre 2025

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

Tecnológico de Monterrey

Introducción

El objetivo de esta práctica fue desarrollar un algoritmo de regresión lineal para entrenar un modelo de aprendizaje automático sin depender de librerías o frameworks especializados. Se trabajó con el dataset *Superstore* (aproximadamente 9,994 registros y 21 columnas), el cual contiene información sobre ventas, descuentos y ganancias de diferentes productos, clientes y regiones. El propósito fue predecir la variable *Profit* (ganancia) mediante dos algoritmos de regresión lineal; el primero con un enfoque sencillo mientras que el segundo se utilizó técnicas de regularización para obtener un mejor resultado.

Metodología

Primero se realizó la preparación de datos, eliminando columnas irrelevantes como identificadores únicos y codificando variables categóricas (categoría, subcategoría, región, segmento y modo de envío). Posteriormente se generaron nuevas variables (features aumentadas) como la diferencia entre la fecha de pedido y la fecha de envío (*DaysToShip*), variables temporales (año, mes y día de la semana) y algunas interacciones entre variables como *Quantity x Discount*.

Después, el dataset se dividió en tres subconjuntos: entrenamiento (60%), validación (20%) y test(20%). Con esta división fue posible ajustar el modelo, evaluar su capacidad de generalización y diagnosticar problemas de sobreajuste o subajuste.

Resultados

El modelo básico de regresión lineal alcanzó un R^2 de aproximadamente 0.28 en el conjunto de prueba, lo cual indica un bajo nivel de ajuste inicial. Sin embargo, al incorporar variables aumentadas y transformaciones, el modelo mejoró de forma considerable, alcanzando un R^2 de 0.69 en el conjunto de prueba. Esto refleja que las nuevas variables aportaron información clave para explicar la variabilidad en las ganancias.

En cuanto al análisis de coeficientes, se encontró que algunas subcategorías como *Copiers* y *Binders* tienen un impacto positivo en la ganancia, mientras que los descuentos elevados, así como productos como *Tables* y *Bookcases*, contribuyen de forma negativa.

Respecto al sesgo y la varianza, el modelo básico presentaba un alto sesgo (bias), lo que significaba que no capturaba adecuadamente la relación entre las variables. Tras la inclusión de nuevas variables, el sesgo disminuyó y la varianza se mantuvo controlada, evitando un sobreajuste excesivo.

Conclusión

La práctica permitió comprender el funcionamiento interno de la regresión lineal y la importancia de la preparación de datos y la ingeniería de variables. Se observó que, aunque un modelo lineal básico puede ser insuficiente para capturar la complejidad de los datos, al mejorar la representación de las variables el desempeño puede incrementarse significativamente. Finalmente, se concluye que la regresión lineal, aun siendo un algoritmo sencillo, puede ofrecer resultados valiosos para la predicción y la interpretación en problemas de negocio reales.