Objetivo del Proyecto

El objetivo fue construir un modelo de regresión para predecir una variable objetivo en un dataset de transacciones financieras. Durante el proceso, se implementaron técnicas de preprocesamiento y se utilizó la regresión lineal como modelo base. La evaluación del modelo se realizó a través de validación cruzada y en un conjunto de prueba separado.

Descripción del Dataset

El dataset contiene columnas como TransactionID, AccountID, TransactionDate, PreviousTransactionDate, TransactionAmount, y variables categóricas y demográficas, como el tipo de transacción y la ocupación del cliente. Las fechas fueron convertidas a formato datetime y luego utilizadas para calcular características de fecha y diferencias temporales, como los días entre transacciones (DaysSincePrevious).

Procesamiento de Datos

Manejo de Valores Nulos: Se utilizaron diferentes métodos para imputar valores nulos según el tipo de columna:

Columnas numéricas: se completaron con la mediana.

Columnas categóricas: se imputaron con el valor más frecuente.

Extracción de Características de Fecha: Las columnas TransactionDate y PreviousTransactionDate fueron utilizadas para extraer año, mes, día y hora de cada transacción, y para calcular DaysSincePrevious, lo cual puede ser relevante para capturar patrones temporales.

División de Datos: El dataset fue dividido en un conjunto de entrenamiento y un conjunto de prueba, permitiendo una evaluación independiente en datos no vistos por el modelo durante el entrenamiento.

Entrenamiento del Modelo y Validación Cruzada

Se utilizó un modelo de regresión lineal como punto de partida para la predicción de la variable objetivo. La evaluación del modelo se realizó a través de validación cruzada y usando el conjunto de prueba:

Métricas de la Validación Cruzada:

MSE (Mean Squared Error) promedio en la validación cruzada 1.013355168521946×10^20 :

Resultados individuales del MSE en cada pliegue:

[4.46e+19, 2.83e+20, 1.09e+20, 3.56e+18, 6.55e+19]

Este MSE promedio muy alto sugiere que el modelo puede estar encontrando dificultades para ajustar los datos de entrenamiento. Podría ser indicativo de alta varianza, outliers o un desequilibrio en los datos que está afectando el rendimiento.

Resultados en el Conjunto de Prueba

Al evaluar el modelo en el conjunto de prueba, se obtuvieron los siguientes resultados:

- Mean Squared Error (MSE): 22,812.77.
- R^2 Score: 0.53.

indica que el modelo logra explicar aproximadamente el 53% de la varianza en los datos de prueba, lo cual es un resultado moderado. Sin embargo, el MSE es significativamente menor en el conjunto de prueba en comparación con los valores de la validación cruzada. Esto sugiere una posible discrepancia en la distribución de los datos entre el conjunto de prueba y los datos de entrenamiento o que algunos valores extremos están afectando los resultados en la validación cruzada.

Análisis de Resultados y Observaciones

Varianza Alta en Validación Cruzada: La diferencia entre el MSE en validación cruzada y en el conjunto de prueba indica que el modelo es sensible a variaciones en los datos. Esto podría ser indicativo de datos no normalizados o con outliers que influyen en los resultados.

Limitaciones del Modelo Lineal: Los datos parecen contener complejidades o no linealidades que el modelo de regresión lineal no está capturando adecuadamente. Esto se refleja en el rendimiento bajo del

Sugerencias para mejoras futuras

Estandarización de los Datos: La estandarización o normalización de los datos podría mejorar la precisión del modelo, especialmente en presencia de valores extremos.

Tratamiento de Outliers: Realizar una identificación y tratamiento de outliers en TransactionAmount y otras variables numéricas podría reducir el impacto de estos en el rendimiento del modelo.

Modelos Alternativos: Probar con modelos más robustos, como RandomForestRegressor o GradientBoostingRegressor, que pueden capturar mejores patrones no lineales en los datos.

Feature Engineering Adicional: Explorar características adicionales, como ratios o interacciones entre variables, podría mejorar el poder predictivo del modelo.

Conclusión

El modelo de regresión lineal ofrece una comprensión inicial, pero los resultados indican que existen factores de complejidad en los datos que este modelo no está capturando bien. Implementar estandarización, gestionar outliers, y evaluar modelos más complejos serían pasos recomendables para mejorar el rendimiento.