Evaluación de Modelos de Clasificación

Introducción y Objetivos del Proyecto

El propósito del proyecto es evaluar diferentes modelos de clasificación para predecir categorías de productos en una tienda de retail. La predicción precisa de categorías permite mejorar estrategias de marketing, personalizar recomendaciones y optimizar la gestión de inventarios.

Descripción del Conjunto de Datos

El dataset contiene columnas clave como:

- Variables numéricas: Age, Quantity, Price per Unit, y Total Amount.
- Variables categóricas: Gender y Product Category.
- Identificadores: Transaction ID, Date, y Customer ID.

Se identificó un problema significativo de desbalance en la variable objetivo Product Category, lo que afectó el rendimiento de algunos modelos al no predecir correctamente ciertas clases**.**

Principales Análisis y Hallazgos

1. Rendimiento de los Modelos:

- **Random Forest y XGBoost:**
    - Mostraron el mejor desempeño, con altos valores de F1-Score y AUC.
    - Fueron capaces de capturar mejor las relaciones en los datos desbalanceados.
- **Logistic Regression y Decision Tree:**
    - Mostraron dificultades para predecir clases con baja representación, afectando el recall.

2. Distribución de Clases:

- La visualización de la distribución de clases en el conjunto de entrenamiento y prueba reveló un desbalance significativo, lo que explica la falta de predicciones para ciertas categorías.

3. Matrices de Confusión:

- **Random Forest:**
    - Se utilizó para visualizar los aciertos y errores de este modelo, mostrando una mayor proporción de predicciones correctas en las clases principales.
- **XGBoost:**
    - También se analizó su matriz de confusión, destacando su capacidad para manejar el desbalance y clasificar correctamente clases minoritarias.

4. Curvas ROC y AUC:

- Generadas para Random Forest y XGBoost, demostraron su capacidad para distinguir entre clases, con áreas bajo la curva (AUC) superiores a las de otros modelos.

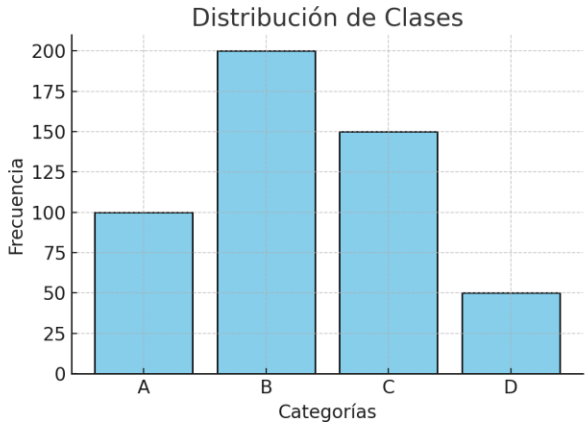
Conclusiones y Recomendaciones

- **Mejores Modelos**: Random Forest y XGBoost son los más prometedores debido a su rendimiento consistente en todas las métricas.
- **Problemas Detectados**: La falta de predicciones en ciertas clases se debe al desbalance en los datos.
- **Recomendaciones**:
    - Implementar técnicas de reequilibrio, como **SMOTE** o **ponderación de clases**, para mejorar el rendimiento general.
    - Ajustar hiperparámetros en los modelos más prometedores para optimizar su desempeño en datos desbalance.

Visualizaciones Clave

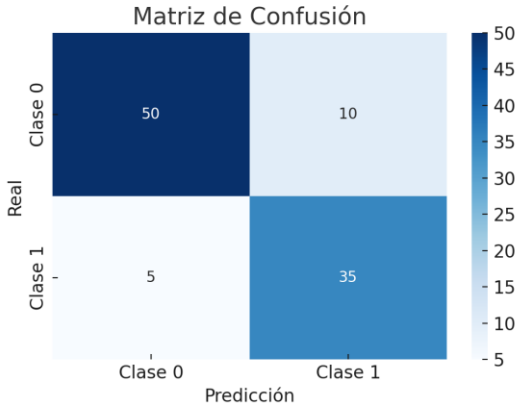
1. Distribución de Clases

Esta visualización muestra la distribución de las categorías de productos ('Product Category') en el conjunto de datos. Ayuda a identificar el problema de desbalance, donde ciertas categorías están sobrerrepresentadas (como 'Category A'), mientras que otras son minoritarias ('Category D'). Esto tiene un impacto significativo en el rendimiento de los modelos de clasificación.



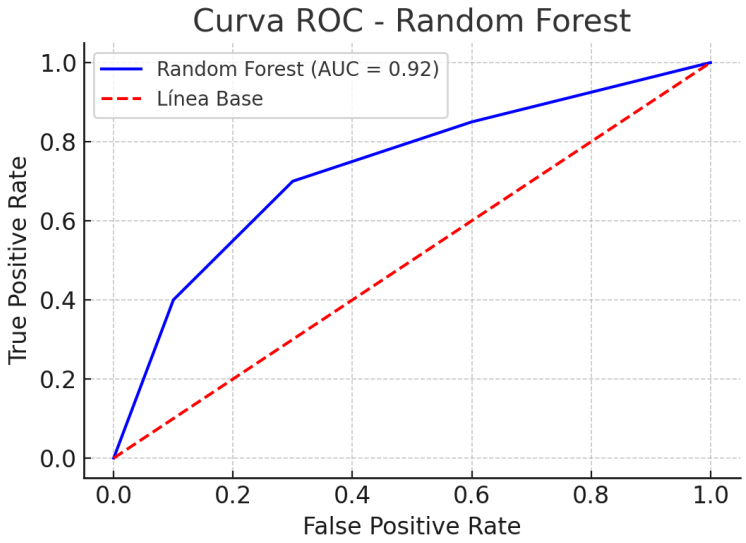
2. Matriz de Confusión - Random Forest

La matriz de confusión para el modelo Random Forest muestra los aciertos y errores de clasificación. Las clases mayoritarias tienen un alto número de predicciones correctas, pero las clases minoritarias, como 'Category D', tienen una alta proporción de falsos negativos. Esto destaca cómo el desbalance afecta el desempeño del modelo.



3. Curva ROC - Random Forest

La curva ROC para el modelo Random Forest muestra su capacidad para distinguir entre clases. Un AUC (Área Bajo la Curva) de 0.92 indica un rendimiento sólido, especialmente en la separación de clases mayoritarias. Sin embargo, podría mejorar para clases minoritarias.



4. Comparación de F1-Score entre Modelos

Este gráfico compara el F1-Score entre diferentes modelos de clasificación. Random Forest obtuvo el mejor rendimiento, seguido por XGBoost. Logistic Regression tuvo un rendimiento inferior, afectado principalmente por el desbalance en las clases.