Análisis de Fraude en Tarjetas de Crédito

1. Introducción al problema

El fraude en tarjetas de crédito representa una amenaza significativa para instituciones financieras y consumidores. Detectar transacciones fraudulentas en tiempo real es un desafío debido al volumen de datos y la baja frecuencia de fraude. Este estudio busca aplicar técnicas de análisis exploratorio y modelado predictivo para identificar patrones de fraude en un conjunto de datos real.

2. Trabajos relacionados

Diversos estudios han abordado el problema del fraude en tarjetas de crédito:
- Dal Pozzolo et al. (2015): 'Calibrating Probability with Undersampling for Unbalanced Classification'.
- Carcillo et al. (2018): 'Combining unsupervised and supervised learning in credit card fraud detection'.
- Bahnsen et al. (2016): 'Cost-sensitive learning with example-dependent costs for credit card fraud detection'.
Estos trabajos destacan la importancia de manejar el desbalance de clases y aplicar técnicas de aprendizaje supervisado y no supervisado.

3. Datos y análisis exploratorio

El dataset utilizado proviene de Kaggle e incluye 284,807 transacciones realizadas por titulares europeos en septiembre de 2013. Contiene 31 columnas, incluyendo variables transformadas mediante PCA (V1 a V28), el monto de la transacción (Amount), el tiempo (Time) y la clase (Class).

Se realizaron las siguientes validaciones:
- No se encontraron valores nulos en columnas ni en filas.
- Se eliminaron 1,081 filas duplicadas.
- La variable 'Class' está altamente desbalanceada: solo 492 transacciones (~0.17%) son fraudulentas.
- Se generaron histogramas para todas las variables y un mapa de calor de correlación para entender relaciones internas.

4. Modelo baseline

Se propone un modelo de regresión logística como línea base para la detección de fraude. Dado el desbalance de clases, se recomienda evaluar el modelo con métricas como precisión, recall, F1-score y AUC-ROC. Además, se sugiere aplicar técnicas de balanceo como SMOTE o undersampling para mejorar el rendimiento del modelo.

5. Conclusiones preliminares

El análisis inicial revela que el dataset está limpio y listo para modelado. La baja proporción de fraudes exige el uso de técnicas especializadas para evitar sesgos en el modelo. Los próximos pasos incluyen la implementación de modelos más robustos, validación cruzada y ajuste de hiperparámetros.