Detección de Fraude en Transacciones Bancarias Usando Aprendizaje Automático

Equipo Parlay

Tecnológico de Monterrey, Campus Guadalajara

Abstract. En este proyecto se desarrolló un sistema de detección de fraude en solicitudes de crédito utilizando técnicas de aprendizaje automático. Se realizó un análisis exploratorio, estrategias de preprocesamiento, modelado con distintos algoritmos, y evaluación del rendimiento. Los resultados muestran los retos de trabajar con datos desbalanceados y las oportunidades de mejora en métricas de recall y F1.

1 Introducción

El fraude financiero representa un problema creciente para las instituciones bancarias, debido a las pérdidas económicas y los riesgos de reputación que conlleva. El objetivo de este proyecto es diseñar, entrenar y evaluar modelos de clasificación para detectar patrones fraudulentos en un conjunto de datos de más de un millón de registros.

2 Metodología

Se trabajó en distintas etapas:

- **Preprocesamiento:** imputación de valores faltantes, escalamiento, detección de outliers y balanceo de clases con técnicas de oversampling.
- **Exploración de datos (EDA):** análisis de distribución de clases, correlaciones, variables atípicas y características relevantes.
- **Modelado:** implementación de regresión logística, árboles de decisión, Random Forest y perceptrón multicapa.
- **Evaluación:** comparación de modelos utilizando métricas como precisión, recall, F1 y matriz de confusión.

3 Experimentos

Se probaron distintos modelos:

- **Regresión logística:** modelo base, sencillo de entrenar pero sensible a outliers y escalamiento.
- **Árboles de decisión y Random Forest:** modelos no lineales que manejan mejor la complejidad y reducen el overfitting.
- **Perceptrón multicapa:** red neuronal con dos capas ocultas para capturar relaciones no lineales.

4 Resultados

Los experimentos muestran que, debido al fuerte desbalance de clases (fraude $< 1\%$), el modelo logra alta exactitud global pero baja precisión en la clase minoritaria. El Random Forest y la red neuronal ofrecieron mejores resultados en comparación con la regresión logística, mejorando el recall y el F1-score.

5 Conclusiones

En conclusión, el proyecto permitió comprobar la dificultad de trabajar con datos altamente desbalanceados en problemas de fraude financiero. Las técnicas de oversampling y modelos más complejos como Random Forest y redes neuronales mejoran el rendimiento, aunque todavía existen limitaciones en la precisión para detectar fraudes. Como equipo, reconocemos la importancia de un preprocesamiento cuidadoso, la selección adecuada de modelos y la necesidad de continuar explorando enfoques más robustos como ensembles avanzados o técnicas de aprendizaje profundo.

6 Bibliografía

References

1. Springer. LNCS – Instructions for Authors. https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines
2. He, H., Garcia, E. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering.
3. Bolton, R., Hand, D. (2002). Statistical Fraud Detection: A Review. Statistical Science.