Ciência da Computação / Sistemas de Informação

Treinando Modelos de Machine Learning para Detecção de Fraudes: Uma Análise de Abordagens com Dados Balanceados e Desbalanceados

Heitor Rodrigues Sabino - 9° módulo de Ciência da Computação, UFLA, bolsista PIBIC/CNPq.

Renato Ribeiro de Lima - Orientador DES, UFLA. - Orientador(a)

Resumo

O aumento exponencial de transações eletrônicas intensificou a incidência de fraudes em cartões de crédito, demandando modelos de detecção cada vez mais precisos. A eficácia de algoritmos de machine learning (ML) para esta tarefa é notória, porém, a performance pode variar drasticamente dependendo da estratégia de tratamento de dados, especialmente em datasets com severo desbalanceamento de classes. Este trabalho compara o desempenho de cinco algoritmos de ML na detecção de fraudes, avaliando o impacto de duas estratégias de treinamento: uma com dados balanceados por sobreamostragem (SMOTE) e outra utilizando os dados em seu estado original desbalanceado. Utilizou-se um dataset público do Kaggle com 284.807 transações (0,17% de fraudes). Os modelos avaliados foram: Random Forest (RF), XGBoost, Regressão Logística (RL), Redes Neurais (NN) e Support Vector Machine (SVM). Todos os modelos tiveram seu limiar de decisão otimizado para maximizar o F1-Score. A performance foi medida por Falsos Negativos/Positivos, Recall, F1-Score e AUC-PR, métricas cruciais para o contexto de fraude. Os resultados indicaram que a melhor estratégia de amostragem é dependente do algoritmo. O modelo RF teve melhor desempenho quando treinado nos dados desbalanceados, alcançando o maior F1-Score geral (0.89) e um excelente Recall (86%). Em contraste, XGBoost (F1-Score igual a 0.86) e RL (F1-Score igual a 0.79) apresentaram desempenho superior quando treinados na base balanceada com SMOTE. Os modelos NN e SVM mostraram performance muito inferior em ambas as estratégias, com o SVM sendo inviabilizado pela necessidade de subamostragem massiva dos dados. A escolha da estratégia de tratamento de dados é um passo crítico e não universal, devendo ser alinhada às características de cada algoritmo. Para este problema, modelos de ensemble como RandomForest e XGBoost provaram ser os mais eficazes. A análise demonstra que o treinamento com dados desbalanceados pode superar técnicas de balanceamento em modelos robustos como o RandomForest, enquanto modelos lineares como a Regressão Logística se beneficiam diretamente do SMOTE.

Palavras-Chave: Detecção de Fraude, Machine Learning, Oversampling.
Instituição de Fomento: CNPQ

Link do pitch: https://www.youtube.com/watch?v=CeO76h66mxI