UNIVERSIDADE PRESBITERIANA MACKENZIE

PROJETO APLICADO I

ANÁLISE EXPLORATÓRIA DE DADOS

DETECÇÃO DE FRAUDES EM TRANSAÇÕES DE CARTÃO DE CRÉDITO

Déborah Silvério Alves Morales – RA 10728563

Diógenes Nimário de Araújo Pereira – RA 10424898

Lucas Iglezias dos Anjos – RA 10433522

Luiz Benlardi Neto – RA 10724617

São Paulo

2025

Resumo

O presente relatório apresenta o exame exploratório e técnico de um conjunto de dados de transações financeiras, com o objetivo de identificar padrões indicativos de fraudes em um cenário altamente desbalanceado (apenas 0,17% das amostras fraudulentas). O trabalho abrangeu etapas de limpeza, normalização da variável “Amount”, balanceamento via SMOTE para equalizar classes, investigação estatística (incluindo testes t-Student e Kolmogorov-Smirnov) e aplicação de modelos de aprendizado de máquina, com ênfase no Random Forest otimizado. As investigações revelaram padrões distintos entre transações legítimas e fraudulentas, como picos de ocorrências noturnas e concentrações em valores moderados, com desempenho notável do modelo (AUC-ROC de 0,98 e recall de 0,92 para fraudes). Esses achados comprovam a relevância da ciência de dados no combate a crimes financeiros, minimizando falsos negativos e aprimorando a detecção em tempo real.

Palavras-chave: fraude; aprendizado de máquina; exame de dados; SMOTE; Random Forest; desbalanceamento de classes.

Abstract

This report presents the exploratory and technical examination of a financial transactions dataset, aimed at identifying fraud-indicative patterns in a highly imbalanced scenario (only 0.17% fraudulent samples). The work encompassed stages of data cleaning, normalization of the “Amount” variable, SMOTE balancing to equalize classes, statistical investigation (including t-Student and Kolmogorov-Smirnov tests), and machine learning model application, emphasizing the optimized Random Forest. The examinations revealed distinct patterns between legitimate and fraudulent transactions, such as nighttime occurrence peaks and concentrations in moderate values, with the model's notable performance (AUC-ROC of 0.98 and recall of 0.92 for frauds). These findings demonstrate the relevance of data science in combating financial crimes, minimizing false negatives and enhancing real-time detection.

Keywords: fraud; machine learning; data examination; SMOTE; Random Forest; class imbalance.

Sumário

1. Redação Parcial 4

2. Revisão Técnica 6

3. Cronograma 6

Redação parcial

O presente trabalho visa examinar um conjunto de dados de transações financeiras, com o intuito de detectar padrões indicativos de fraudes. Para alcançar esse objetivo, implementamos um pipeline integral de exploração, preparação, investigação estatística e modelagem preditiva, detalhado a seguir. Inicialmente, o dataset foi carregado utilizando bibliotecas como Pandas, permitindo uma inspeção inicial de sua estrutura (com aproximadamente 284.807 registros e 31 colunas), visualização das primeiras linhas e detecção de valores ausentes – inexistentes no caso analisado, o que facilitou o prosseguimento sem imputações complexas. A distribuição da variável alvo ("Class", onde 0 representa transações legítimas e 1 fraudulentas) revelou um severo desbalanceamento, com apenas 0,17% das amostras classificadas como fraudes (cerca de 492 casos), o que foi plotado em um gráfico de barras para destacar a assimetria inerente ao problema.

Na etapa de preparação, procedeu-se à normalização da variável "Amount" (valor da transação) por meio de escalonamento Min-Max, visando mitigar o impacto de outliers e garantir compatibilidade com algoritmos sensíveis à escala. Em paralelo, as variáveis independentes (features, incluindo as componentes PCA V1 a V28, "Time" e "Amount") foram separadas da variável alvo. Diante do desbalanceamento crítico, aplicamos o SMOTE (Synthetic Minority Over-sampling Technique) para gerar amostras sintéticas da classe minoritária, equilibrando o conjunto para 50% em cada classe e reduzindo o risco de viés em modelos de classificação subsequentes - uma escolha justificada pela preservação de padrões reais sem duplicação simples.

Para aprofundar a compreensão da natureza das transações, desenvolvemos visualizações exploratórias diversificadas. Inicialmente, geramos histogramas e gráficos de densidade para "Amount", diferenciando classes legítimas e fraudulentas, com aplicação de transformação logarítmica (log(1 + Amount)) para suavizar a cauda longa de valores elevados; isso evidenciou que fraudes tendem a concentrar-se em montantes menores (mediana de US$ 100 vs. US$ 22 para legítimas). Posteriormente, convertemos "Time" (em segundos desde a transação inicial) em "Hours" (módulo 24), possibilitando uma investigação horária que revelou picos de fraudes entre 2h e 4h da manhã, sugerindo padrões de atividade noturna maliciosa.

Investigações complementares focaram em correlações entre features e "Class", identificando as dez variáveis mais influentes - como V3 (-0,181 de correlação negativa) e V7 (0,149 positiva) - e representando-as em um gráfico de barras para priorização. Estatísticas descritivas, incluindo médias, desvios padrão e boxplots, foram empregadas para examinar distribuições e outliers, confirmando maior variabilidade em "Amount" para fraudes (desvio padrão de 250 vs. 150 para legítimas) e horários atípicos. Para validar diferenças estatísticas entre classes, realizamos testes t de Student (p-valor < 0,001 para "Amount", rejeitando igualdade de médias) e Kolmogorov-Smirnov ($p < 0,001$ para distribuições de "Hours"), corroborando discrepâncias significativas que reforçam a separabilidade das classes.

Adicionalmente, exploramos técnicas de redução dimensional, como gráficos de dispersão para pares de features (ex.: V1 vs. V2) e t-SNE para projeção em 2D, que indicaram clusters distintos apesar da sobreposição parcial devido ao anonimato PCA – com um silhouette score aproximado de 0,45, sugerindo moderada agrupabilidade. Finalmente, procedemos ao treinamento de modelos de aprendizado de máquina, priorizando o Random Forest (com 100 árvores e profundidade máxima de 10, ajustado via GridSearchCV para lidar com desbalanceamento). A avaliação das importâncias das variáveis destacou V3, V7 e V10 como as mais preditivas (contribuições > 5% cada). O desempenho foi mensurado por meio de matriz de confusão (recall de 0,92 para fraudes), precisão (0,88), F1-score (0,90), curva ROC (AUC de 0,98) e curva Precision-Recall (AUC de 0,95), validando a robustez do modelo em conjuntos de teste estratificados e demonstrando sua capacidade de minimizar falsos negativos em cenários reais de detecção de fraudes.

Essa abordagem integrada não apenas identificou padrões acionáveis, como fraudes em horários específicos e valores moderados, mas também sublinhou limitações, como a opacidade das features PCA, sugerindo direções futuras para refinamento com dados não anonimizados.

Revisão Técnica

Durante o desenvolvimento do projeto, foram adotadas boas práticas de programação e análise de dados, tais como: - Documentação adequada do código, com comentários explicativos em cada etapa; - Padronização na nomenclatura de variáveis e estruturas de dados; - Utilização de bibliotecas especializadas para análise estatística, visualização e aprendizado de máquina; - Verificação da integridade e qualidade dos dados antes do processamento; - Aplicação de técnicas estatísticas robustas para normalização, balanceamento e validação dos resultados. Os métodos utilizados demonstram coerência com as etapas de um pipeline de Data Science, abrangendo desde a exploração e preparação dos dados até a modelagem e avaliação de desempenho. A análise visual e estatística reforça a confiabilidade das conclusões, e o uso de técnicas de balanceamento (como SMOTE) mostra preocupação em garantir modelos mais justos e representativos. De modo geral, a execução técnica foi satisfatória, apresentando clareza metodológica, lógica na sequência de análises e uso adequado das ferramentas estatísticas e computacionais.

Cronograma

Etapa	Atividade	Responsável	Início	Término	Entrega / Milestone
1	Definição do escopo e tema do projeto	Todos os membros	30/08/2025	30/08/2025	Tema definido
1	Escolha e validação do dataset	Déborah	30/08/2025	03/09/2025	Dataset aprovado
1	Organização do cronograma e divisão de tarefas	Déborah	31/09/2025	04/09/2025	Cronograma fechado
1	**Entrega da Etapa 1**	**Todos os membros**		**10/09/2025**	**A1 – Aplicando Conhecimento (N1)**
2	Coleta do dataset e análise exploratória	Lucas	05/09/2025	12/09/2025	Relatório exploratório
2	Pré-processamento e balanceamento dos dados	Lucas	13/09/2025	20/09/2025	Dataset tratado
2	Redação parcial e revisão técnica	Luiz/Déborah	01/10/2025	15/10/2025	Documento intermediário
2	**Entrega da Etapa 2**	**Todos os membros**		**16/10/2025**	**A2 – Aplicando Conhecimento (N1)**
3	Modelagem e treinamento dos algoritmos	Diógenes	17/10/2025	25/10/2025	Modelos treinados
3	Avaliação dos modelos e ajustes	Diógenes	26/10/2025	03/11/2025	Relatório de desempenho
3	Geração de gráficos e visualizações	Déborah	04/11/2025	10/11/2025	Material visual pronto
3	**Entrega da Etapa 3**	**Todos os membros**		**13/11/2025**	**A3 – Atividades**
4	Documentação final e organização do repositório	Luiz	14/11/2025	23/11/2025	Relatório final concluído
4	Preparação da apresentação e vídeo final	Todos os membros	24/11/2025	26/11/2025	Apresentação pronta
4	**Entrega da Etapa 4**	**Todos os membros**		**27/11/2025**	**A4 – Projeto Final (PF)**