

DETECÇÃO DE FRAUDES FINANCEIRAS COM APRENDIZADO DE MÁQUINA: O IMPACTO DO BALANCEAMENTO DE CLASSES

Jairo de Cássio Teixeira Filho¹
Timoteo Augusto Soares Teixeira²
Roselio Marcos Santana³

timoteoasteixeira@gmail.com

ÁREA DO CONHECIMENTO: Ciências Sociais e Aplicadas

PALAVRAS-CHAVE: fraudes financeiras; aprendizado de máquina; balanceamento de classes.

1 INTRODUÇÃO

O avanço das transações financeiras digitais impulsionou a eficiência dos ecossistemas de pagamento, mas expandiu a superfície de vulnerabilidade para fraudes eletrônicas. O enfrentamento dessas ameaças exige a adoção de sistemas de segurança automatizados e de alta precisão. Nesse cenário, o uso de algoritmos de Aprendizado de Máquina (*Machine Learning*) consolidou-se como uma abordagem técnica indispensável para a detecção de padrões anômalos em grandes volumes de transações. Apesar da eficácia computacional desses algoritmos, a detecção de fraudes enfrenta um obstáculo inerente à natureza dos dados financeiros: o severo desbalanceamento de classes. Em cenários reais, as transações fraudulentas representam uma fração mínima em comparação às transações legítimas. Modelos preditivos submetidos a bases de dados altamente assimétricas tendem a focar na classe majoritária, reduzindo drasticamente a capacidade de identificar fraudes, o que resulta em um elevado índice de falsos negativos. A literatura acadêmica aponta o uso de técnicas de balanceamento de dados e penalizações algorítmicas como soluções, contudo, estudos indicam uma lacuna metodológica na comparação sistemática do impacto dessas técnicas quando aplicadas a diferentes arquiteturas de modelos, especialmente considerando o equilíbrio entre o desempenho preditivo sobre a classe minoritária e o custo computacional em cenários de alta dimensionalidade. Diante dessa lacuna, emerge o seguinte problema de pesquisa: como a aplicação de diferentes técnicas de tratamento de dados desbalanceados impacta a capacidade preditiva e a viabilidade operacional de modelos de aprendizado de máquina na detecção de fraudes financeiras? Como hipótese central, postula-se que técnicas de superamostragem sintética apresentarão a maior eficácia na mitigação de falsos negativos, porém ao custo de uma latência

¹ Acadêmico do 7º período de Ciência da Computação – Centro Universitário Vértice – Univértix – Matipó.

² Acadêmico do 7º período de Ciência da Computação – Centro Universitário Vértice – Univértix – Matipó.

³ Professor do curso de Ciência da Computação - Centro Universitário Vértice – Univértix – Matipó.

computacional elevada, enquanto o ajuste nativo de pesos algorítmicos entregará a melhor relação de custo- benefício operacional. Para responder a essa problemática, este trabalho estabelece como objetivo geral investigar o impacto de técnicas de balanceamento de classes na eficácia preditiva e no custo computacional de diferentes modelos de aprendizado de máquina aplicados à detecção de fraudes em cartões de crédito. Para o alcance deste propósito, definem-se os seguintes objetivos específicos: implementar os algoritmos preditivos de Regressão Logística, *Random Forest* e *XGBoost* em um ambiente de simulação; aplicar técnicas de balanceamento em nível de dados e em nível de algoritmo; avaliar o desempenho dos modelos por meio de métricas apropriadas para dados assimétricos e o tempo de processamento computacional; e analisar estatisticamente as variações de desempenho entre os cenários propostos para determinar as abordagens mais eficientes. A pesquisa justifica-se cientificamente por fornecer um referencial empírico rigoroso sobre o comportamento de algoritmos sob forte assimetria de classes. Do ponto de vista prático, o estudo oferece embasamento técnico para que instituições financeiras parametrizem sistemas de segurança que equilibrem a detecção de anomalias com a restrição de latência exigida em ambientes de produção.

2 METODOLOGIA

Este trabalho classifica-se como uma pesquisa quantitativa de natureza empírico-experimental. A abordagem fundamenta-se na manipulação de variáveis independentes, representadas pelas técnicas de balanceamento, para observar seus efeitos sobre as variáveis dependentes, que compreendem o desempenho preditivo e o custo computacional, utilizando dados secundários em ambiente controlado.

Contexto e Delineamento Experimental A base de dados adotada será o conjunto de dados de detecção de fraudes em cartões de crédito disponível publicamente. O arquivo contempla 284.807 transações, com apenas 0,172% de fraudes. As variáveis compreendem 28 componentes numéricos resultantes de uma transformação por análise de componentes principais, além das características originais de tempo e valor da transação. Os experimentos serão conduzidos em linguagem Python, utilizando as bibliotecas *scikit-learn* e *imbalanced-learn*. A fim de garantir a validade científica das métricas de custo computacional, o experimento será executado em um ambiente computacional padronizado, cujas especificações de hardware serão rigorosamente documentadas na etapa de apresentação dos resultados. No pré- processamento, a variável de tempo será normalizada. Para a variável de valor, visando mitigar o impacto de valores atípicos típicos de transações financeiras, será aplicada a técnica de escalonamento robusto, que utiliza a mediana e a amplitude interquartil. A semente aleatória global será fixada em 42, e a validação seguirá a abordagem de validação cruzada estratificada com cinco partições, garantindo a manutenção da proporção da classe alvo em todos os testes. O núcleo preditivo testará os algoritmos de Regressão Logística, *Random Forest* e *XGBoost*. Os modelos serão inicializados utilizando seus hiperparâmetros padrão, garantindo que as variações observadas decorram estritamente das técnicas de balanceamento aplicadas. O estudo envolverá quatro cenários analíticos: o primeiro atuará como um cenário de controle, realizando o treinamento sem mitigação; o segundo aplicará a redução aleatória da classe majoritária; o terceiro adotará o SMOTE para a superamostragem da classe

minoritária, mediante monitoramento do consumo de memória dada a expansão da matriz de dados; e o quarto aplicará o balanceamento de custos nativo dos algoritmos, utilizando os parâmetros específicos de peso para o ajuste matemático diretamente na função de perda.

O desempenho será mensurado pelas métricas de precisão, revocação, F1- Score e as áreas sob as curvas ROC e de precisão-revocação. O custo computacional será extraído através do tempo de execução em segundos das funções de ajuste e predição. Para validar se as variações de desempenho são estatisticamente significativas, será aplicado o teste de postos sinalizados de Wilcoxon com um nível de significância de 5%. O estudo dispensa aprovação de comitê de ética por utilizar base de domínio público plenamente anonimizada.

3 RESULTADOS E DISCUSSÃO

O presente estudo encontra-se em fase de desenvolvimento, por se tratar de um Trabalho de Conclusão de Curso (TCC). Atualmente, a pesquisa aloca-se na etapa de levantamento bibliográfico e construção da API Smart Stock, que atua como base preparatória para a execução dos testes. Neste estágio, as principais referências teóricas e metodológicas já foram sistematizadas. A partir disso, evidenciou-se que as micro e pequenas empresas (MPEs) enfrentam altos índices de mortalidade precoce decorrentes, em parte, da gestão deficiente de estoque e da baixa maturidade digital (SEBRAE, 2024; CGI.br, 2024). A aplicação dos princípios SOLID e da arquitetura em camadas, combinados ao ecossistema Java/Spring Boot, surge como resposta para garantir sistemas coesos, escaláveis e acessíveis (Santos; Lima, 2023; Costa et al., 2024). As etapas subsequentes compreendem a finalização do desenvolvimento da API, a execução dos testes com dados sintéticos e a análise dos resultados, previstas para o segundo semestre de 2026. Estima-se a conclusão e formatação final do trabalho em outubro de 2026.

4 CONSIDERAÇÕES FINAIS

Por se tratar de um TCC em desenvolvimento, o parecer final será apresentado após a finalização da pesquisa. Espera-se que este estudo demonstre a viabilidade técnica do sistema Smart Stock como uma solução escalável e acessível para a gestão de estoque de MPEs. Além disso, a pesquisa pretende comprovar que a aplicação dos princípios SOLID e da arquitetura em camadas, aliada ao ecossistema Java, Spring Boot e PostgreSQL, resulta em um sistema coeso, de alta usabilidade e baixo custo de manutenção, contribuindo para a digitalização e fortalecimento da competitividade dos pequenos negócios brasileiros.

3 RESULTADOS E DISCUSSÃO

O presente estudo encontra-se em fase de desenvolvimento, por se tratar de um Trabalho de Conclusão de Curso (TCC). Atualmente, a pesquisa aloca-se na etapa de levantamento bibliográfico.

4 CONSIDERAÇÕES FINAIS

Por se tratar de um TCC em desenvolvimento, o parecer final será apresentado após a finalização da pesquisa

REFERÊNCIAS

AHMAD, H. *et al.* Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS). **International Journal of Information Technology**, v. 15, n. 1, p. 325-333, 2023.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CHAWLA, N. V. *et al.* SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321-357, 2002.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. 2016. p. 785-794.

HERATH, H. M. M. N. Advancing Machine Learning for Financial Fraud Detection: A Comprehensive Review of Algorithms, Challenges, and Future Directions. **ASEAN Journal of Economic and Economic Education**, v. 4, n. 1, p. 49-68, 2025.

IMANI, M. *et al.* Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers. **Technologies**, v. 14, n. 1, p. 54, 2026.

THEODORAKOPOULOS, L. *et al.* Big Data-Driven Distributed Machine Learning for Scalable Credit Card Fraud Detection Using PySpark, XGBoost, and CatBoost. **Electronics**, v. 14, n. 9, p. 1754, 2025.