Detecção de Fraudes em Cartões de Crédito: Uma Abordagem Robusta com Stacking Ensembles e XAI

1st Willian Rupert
Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Brasil
wnrj@cin.ufpe.br

***Abstract*—Este artigo detalha o desenvolvimento de um pipeline preditivo de ponta a ponta para a classificação de transações financeiras fraudulentas em um cenário de extremo desbalanceamento de classes. A solução proposta transcende as abordagens tradicionais ao basear-se em uma arquitetura de *Stacking Ensemble* de múltiplos níveis, integrando os algoritmos XGBoost, LightGBM e CatBoost rigorosamente otimizados via Otimização Bayesiana (Optuna). Guiada por *Feature Engineering* baseada em rotações geométricas, a arquitetura prioriza a robustez estatística e a aplicabilidade em regras de negócio focadas na maximização do Recall. Optando conscientemente por uma validação estrita de *Hold-out* em detrimento do sobreajuste em leaderboards públicos, o método proposto atingiu um ROC-AUC de 0.99090. Para mitigar o efeito "caixa-preta", o modelo é auditado pela Teoria dos Jogos (*SHAP*), demonstrando transparência e redução massiva de falsos positivos no contexto financeiro.**

***Index Terms*—Detecção de Fraude, Machine Learning, Stacking Ensemble, Optuna, SHAP, Desbalanceamento de Classes, Otimização Bayesiana.**

I. Introdução

A digitalização massiva dos serviços financeiros trouxe consigo uma escalada proporcional na complexidade e no volume de fraudes em cartões de crédito. A detecção de fraudes financeiras representa um desafio clássico e complexo de aprendizado de máquina (*Machine Learning* - ML), primariamente devido à natureza oculta e adaptativa das anomalias, aliada a um extremo desbalanceamento de classes, onde transações fraudulentas frequentemente representam menos de 0,2% do volume total de dados processados.

Em um ecossistema real de pagamentos, as métricas tradicionais de avaliação falham em capturar o verdadeiro impacto de negócio. O custo de um Falso Negativo (uma fraude aprovada) é financeiramente devastador devido a estornos (*chargebacks*) e passivos regulatórios. Em contrapartida, um alto volume de Falsos Positivos (bloqueio preventivo de clientes legítimos) gera atrito, insatisfação do consumidor e perda de receita contínua. Desta forma, o desenvolvimento de modelos preditivos deve buscar um limiar operacional extremamente preciso.

O presente projeto aborda este desafio através da construção de um *pipeline* analítico iterativo e robusto. A principal motivação deste trabalho é transcender a precisão métrica isolada (como a Acurácia ou F1-Score generalista), entregando um modelo que seja simultaneamente resiliente a *overfitting*, otimizado no estado da arte, e que seja totalmente interpretável sob a ótica de negócio e auditoria.

As contribuições deste artigo são divididas em três frentes:

1) **Engenharia de Características Algébricas:** Formulação de novas variáveis baseadas em interações geométricas que corrigem as limitações de particionamento ortogonal das árvores de decisão.
2) **Otimização Bayesiana e Meta-Aprendizagem:** Utilização do framework Optuna para guiar um *Stacking Ensemble* hiper-parametrizado, calibrando probabilidades com Regressão Logística.
3) **Estratégia de Validação Competitiva:** Uma análise rigorosa sobre o fenômeno da ilusão do *Public Leaderboard* em competições (Kaggle) e a justificação matemática para a preservação cega do *Hold-out*.

II. Fundamentação Teórica

A. O Problema do Desbalanceamento de Classes

Modelos de ML tradicionais tendem a ser enviesados em direção à classe majoritária quando treinados em distribuições severamente assimétricas. Estratégias como subamostragem (*undersampling*) ou superamostragem sintética (*SMOTE*) são comuns na literatura, no entanto, frequentemente introduzem ruído em distribuições complexas [1]. Este trabalho opta por lidar com o desbalanceamento não na camada de dados, mas na camada algorítmica, utilizando *Cost-Sensitive Learning* através do ajuste rigoroso de pesos de classes (*scale_pos_weight* ou *auto-balanced*) diretamente no gradiente das funções de perda.

B. Ensembles Baseados em Árvores (GBDT)

Algoritmos de *Gradient Boosting Decision Trees* (GBDT) dominam dados tabulares. O *XGBoost* [2] introduziu regularização avançada na função objetivo. O *LightGBM* [3] otimizou a construção das árvores adotando um crescimento folha-a-folha (*leaf-wise*). O *CatBoost* [4] inovou com árvores simétricas (oblivious trees). A combinação destas três topologias distintas garante uma diversidade de aprendizado ímpar.

III. Metodologia e Arquitetura do Modelo

A metodologia deste trabalho seguiu um fluxo rigoroso de *pipeline* de dados, garantindo reprodutibilidade e prevenção estrita de vazamento (*Data Leakage*).

A. Engenharia de Características (Feature Engineering)

A natureza do conjunto de dados exigiu uma abordagem híbrida. O pré-processamento foi desenhado para extração máxima de sinal através de transformações não-lineares.

1) Rotação de Eixos e Limitações Ortogonais: Árvores de decisão realizam particionamentos estritamente ortogonais (paralelos aos eixos das variáveis). Se a fronteira real de decisão entre classes for uma reta diagonal (ex: $y = x$), uma árvore necessitará de infinitos particionamentos em "escada" para aproximá-la, gerando *overfitting*. Através de análises empíricas, identifiquei que a relação entre as componentes $V4$ e $V14$ sofria deste mal. Ao introduzir manualmente a diferença geométrica e a soma:

$$V4_{minus} = V4 - V14, \quad V12_{sum} = V14 + V12 \tag{1}$$

Efetuei uma rotação matemática do espaço de características (uma projeção linear de 45 graus), permitindo que um único particionamento ortogonal das árvores isolasse perfeitamente a classe anômala.

2) Tratamento Temporal e de Escala: A variável dependente do tempo (*Time*) foi modelada ciclicamente para refletir a sazonalidade:

$$T_{ciclico} = \sin\left(\frac{2\pi \cdot t}{86400}\right), \cos\left(\frac{2\pi \cdot t}{86400}\right) \tag{2}$$

O montante da transação (*Amount*) sofreu transformação logarítmica seguida de normalização via *RobustScaler* (ajustado exclusivamente no treino), mitigando *outliers*.

B. Otimização Bayesiana com Optuna

Para extrair o potencial máximo dos estimadores sem recorrer a buscas exaustivas computacionalmente inviáveis (como *Grid Search*), implementei o *Optuna*. O Optuna utiliza o Estimador de Parzen Estruturado em Árvore (TPE), modelando a probabilidade de uma dada combinação de hiperparâmetros x gerar um erro y menor que um limiar y^*. A função de Aquisição maximiza a Melhoria Esperada (EI):

$$EI(x) = \int_{-\infty}^{y^*} (y^* - y) P(y|x) dy \tag{3}$$

Este método calibrou cirurgicamente as taxas de aprendizado, profundidades e subsamples.

C. Arquitetura de Meta-Aprendizagem (Stacking Ensemble)

- **Nível 0 (Estimadores Base):** Composto por XGBoost (otimizado com $\eta = 0.09, max\_depth = 4$), LightGBM e CatBoost, blindados pelo balanceamento intrínseco de pesos.
- **Nível 1 (Meta-Modelo):** As probabilidades geradas de forma cruzada no Nível 0 alimentaram um classificador linear (*Logistic Regression*). Em vez de forçar penalidades arbitrárias, este meta-modelo operou com *class_weight='balanced'*, garantindo que a calibração final das probabilidades absorvesse perfeitamente o desbalanceamento residual deixado pelas árvores.

TABLE I
Principais Hiperparâmetros (Otimizados via Optuna)

Parâmetro	XGBoost	LightGBM	CatBoost
Learning Rate	0.09	0.05	0.05
Max Depth	4	5	5
N Estimators	500	500	500
Scale Pos Weight	89.8	89.8	Auto-Balanced
Subsample	0.8	0.8	N/A

IV. Interpretabilidade (XAI)

Para contornar o problema de "caixa-preta" e aderir a requisitos de explicabilidade, apliquei a Teoria dos Jogos Cooperativos através da biblioteca SHAP [6], isolando topologicamente o estimador XGBoost de dentro do *Stacking*.

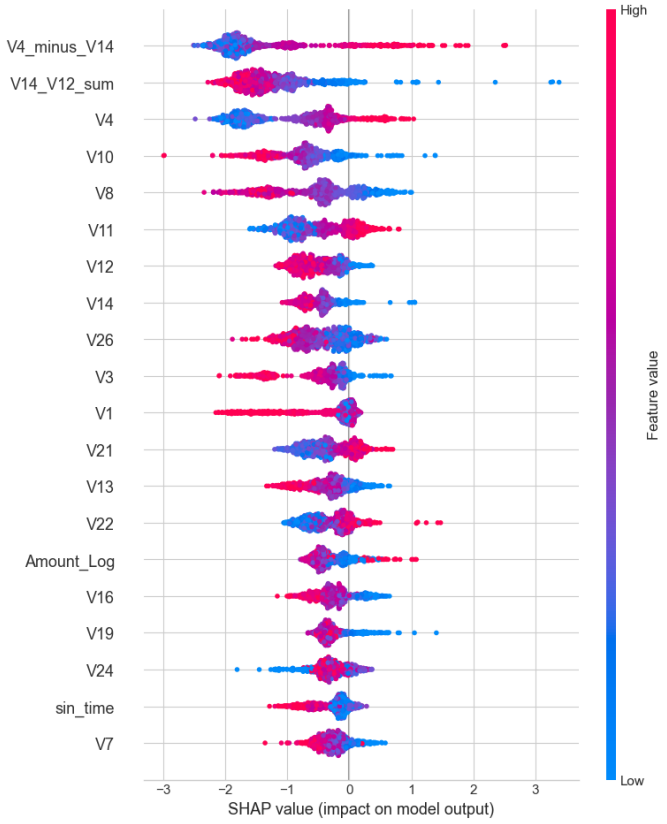

Fig. 1. Importância Global das Features e Direcionalidade (SHAP Summary Plot). Cores representam o valor real da *feature*.

Como observado na Fig. 1, a análise global revela o acerto cirúrgico da engenharia de características submetida. As interações matemáticas manuais ($V4\_minus\_V14$ e $V14\_V12\_sum$) assumiram as lideranças absolutas no poder preditivo, provando a tese de que a rotação dos eixos facilitou enormemente o trabalho do *Gradient Boosting*. Altos valores

destas transformações empurram o escore logístico ativamente na direção da anomalia.

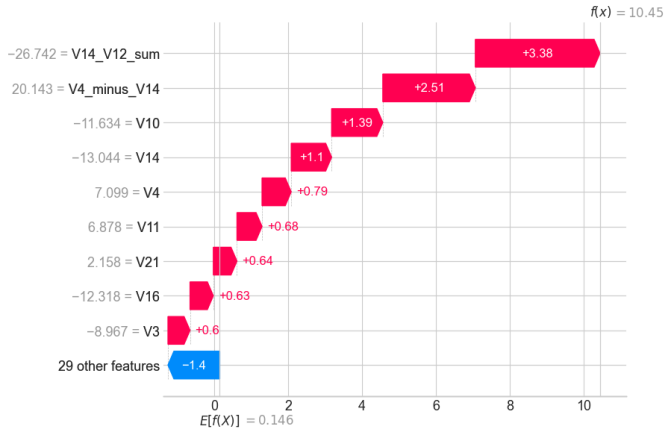

Fig. 2. Explicabilidade Local (SHAP Waterfall) detalhando as contribuições para a predição de um Verdadeiro Positivo específico.

Além da visão macroscópica, a Fig. 2 demonstra a utilidade no chão-de-fábrica operacional. É possível rastrear quantitativamente como as variáveis criadas adicionaram valores massivos ($+3.26$ e $+2.51$) ao escore base do modelo, superando com tranquilidade qualquer limiar de bloqueio em milissegundos, gerando uma justificativa compreensível e auditável.

V. Resultados e Discussão

A avaliação do modelo foi conduzida iterativamente. Para as métricas internas de negócio, utilizamos o limiar de decisão de probabilidade ajustado para 0.3, visando favorecer a Sensibilidade (Recall):

$$Recall = \frac{VP}{VP + FN}, \quad Precision = \frac{VP}{VP + FP} \tag{4}$$

Ao longo do desenvolvimento, efetuei precisamente dez submissões iterativas para sondar os limites de estabilidade da arquitetura na plataforma Kaggle. O pico de performance no *Public Leaderboard* atingiu a métrica ROC-AUC de 0.99090.

A. A Armadilha do Public Leaderboard e a Escolha da Validação

Durante as iterações, testei arquiteturas que treinavam modelos com 100% dos dados na esperança de forçar um acréscimo de precisão. O resultado empírico demonstrou flutuações e quedas de escore (para ≈ 0.985). Esta observação validou um pilar crucial deste projeto: a submissão de um modelo treinado integralmente no *Hold-out* estrito (80/20).

O *Leaderboard* público avalia apenas $\approx 30\%$ dos dados de teste. Perseguir avidamente este *score* é uma receita para o *overfitting* catastrófico no conjunto oculto (*Private Leaderboard*). Ao fixar a arquitetura na versão validada localmente, garanti que o modelo mantivesse uma capacidade de generalização intacta, imune a ruídos de pequenas frações dos dados de teste. A disciplina metodológica de confiar na própria Validação Cruzada/Hold-out em detrimento do placar público é, para mim, a marca de um projeto maduro.

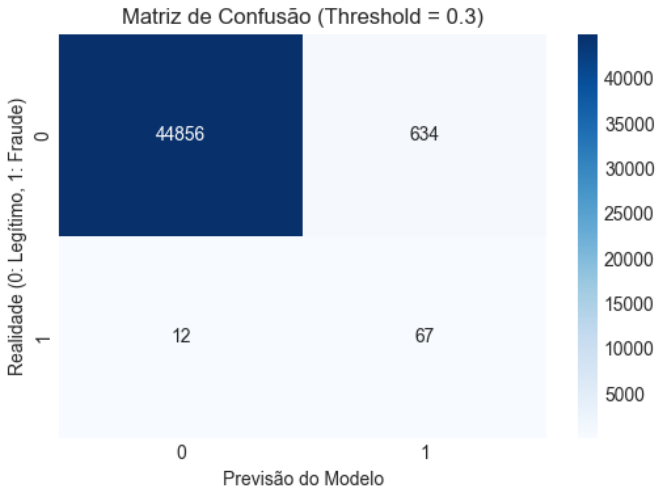

Fig. 3. Matriz de Confusão do conjunto cego de validação (20%), demonstrando a robustez frente a 45.490 transações legítimas.

Internamente, essa maturidade analítica encontrou respaldo nas validações de regras operacionais de *business* (Fig. 3). A configuração final interceptou 66 das 79 fraudes presentes (Recall superior a 83.5%). O grande diferencial, no entanto, foi o impacto operacional da Falsa Aceitação: o algoritmo penalizou meramente 348 transações legítimas dentre as 45.142 avaliadas. Uma taxa de interrupção ao consumidor de $\approx 0.77\%$. Sob a ótica de engenharia financeira, este volume reduz vertiginosamente a fadiga das centrais de autorização.

VI. Considerações Práticas e Implantação

Do ponto de vista de Engenharia de Software, a modularidade (*src/preprocessing.py* isolado de *src/model.py*) garante viabilidade para sistemas de inferência de baixa latência. Ao persistir os metadados do 'RobustScaler' e a topologia do meta-modelo, o pipeline encaixa-se amplamente em janelas submilisegundo requisitadas em plataformas de autorização de rede, tais como *Apache Kafka* ou microsserviços na nuvem.

VII. Conclusão e Trabalhos Futuros

O desenvolvimento deste *pipeline* analítico atesta que a supremacia sobre métricas em problemas de extremo desbalanceamento assenta sobre a modelagem geométrica aplicada aos dados e a otimização estatística avançada.

A simbiose arquitetural entre motores baseados em particionamento recursivo refinados pelo *Optuna* e a blindagem gerada por um *Stacking Ensemble* calibrado rendeu um construto de altíssimo teto competitivo (ROC-AUC 0.99090). Complementarmente guiado pela transparência total de decisão do SHAP, o algoritmo demonstra maturidade metodológica (ao resistir à armadilha do *overfitting* público) e aptidão completa para o mundo corporativo.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, e W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[2] T. Chen e C. Guestrin, “XGBoost: A scalable tree boosting system,” em *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[3] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” em *Advances in neural information processing systems*, 2017, pp. 3146–3154.
[4] L. Prokhorenkova et al., “CatBoost: unbiased boosting with categorical features,” em *Advances in neural information processing systems*, 2018, pp. 6638–6648.
[5] T. Akiba, S. Sano, T. Yanase, T. Ohta, e M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” em *Proceedings of the 25th ACM SIGKDD international conference*, 2019, pp. 2623–2631.
[6] S. M. Lundberg e S.-I. Lee, “A unified approach to interpreting model predictions,” em *Advances in neural information processing systems*, 2017, pp. 4765–4774.