

UPKF Scientific Draft

Title: Deteccao de Fraudes em Cartoes com Redes Neurais

Category: research

Type: ScholarlyArticle

Year: 2025

Author: Carlos Ulisses Flores

Resumo

Este trabalho apresenta um estudo de caso de deteccao de fraudes em cartoes de credito sob forte desbalanceamento, comparando um Perceptron Multi-Camadas (MLP) supervisionado as alternativas Autoencoder (AE), Regressao Logistica (LR) e Isolation Forest (IF) no conjunto publico ULB/Worldline. O protocolo prioriza metricas apropriadas a classes raras, em especial AUC-PR e F1 (alem de F), com thresholds calibrados na validacao e aplicados no teste; reportamos curvas ROC/PR, matrizes de confusao, importancia por permutacao e teste de robustez a variacoes de prevalencia. O MLP obteve o melhor F1 na classe positiva e AUC-PR competitiva, superando AE/IF e empatando/superando LR; discutimos escolha de limiar sensivel a custos, calibracao e governanca, com artefatos completos para replicacao (SAITO; REHMSMEIER, 2015; DAVIS; GOADRICH, 2006).

1. Introducao

O objetivo deste trabalho e aplicar tecnicas de Deep Learning, especificamente um Perceptron Multi-Camadas (MLP), para melhorar a deteccao de fraudes em cartoes de credito, comparando-o com abordagens alternativas como Autoencoder, Regressao Logistica e Isolation Forest, em um contexto de forte desbalanceamento de classes, utilizando o conjunto de dados publico ULB/Worldline. Fraude em meios electronicos exige maximizar a sensibilidade sob baixas taxas de falsos positivos, alinhando risco operacional a precisao estatistica, sobretudo quando a taxa base de fraude e infima. Este cenario motiva metricas mais informativas do que acuracia bruta. A avaliacao deve refletir custo assimetrico $FN > FP$ (HERNANDEZ AROS et al., 2024; CHERIF et al., 2023; CHEN et al., 2024). Em classes raras, curvas PR e medidas F1/F descrevem melhor o trade-off entre precisao e revocacao do que ROC isolada. A leitura de PR evita superestimar desempenho por negativos abundantes (SAITO; REHMSMEIER, 2015; DAVIS; GOADRICH, 2006). Estabelecemos baselines fortes (LR, IF), comparamos com AE nao supervisionado e posicionamos MLP como candidato principal em pipeline reproduzivel (KING; ZENG, 2001; LIU; TING; ZHOU, 2008).

Fundamentacao Teorica

Regressao Logistica e referencia classica para eventos raros, fornecendo probabilidades calibraveis e fronteira linear interpretavel. Com regularizacao adequada, evita sobreajuste e mantem estabilidade (KING; ZENG, 2001; NICULESCU-MIZIL; CARUANA, 2005). Detectores nao supervisionados capturam anomalias estruturais, porem falham quando fraudes sao quase normais no espaco latente. Erro de reconstrucao pode produzir muitos falsos positivos (LIU; TING; ZHOU, 2008; DAL POZZOLO, 2015). Calibracao de probabilidades (Platt/Temperature) impacta governanca ao alinhar escores a probabilidades bem-comportadas (NICULESCU-MIZIL; CARUANA, 2005; GUO et al., 2017). A literatura recente de desbalanceamento reforca PR-AUC como metrica primaria e sugere protocolos especificos de avaliacao (CHEN et al., 2024).

Dados e Preparacao

Utilizamos ULB/Worldline (284.807 transacoes; 30 variaveis PCA + Time/Amount; Class

binario), cenario tipico de classe rara com 0,173% de fraudes (ULB/Worldline, 2013; DAL POZZOLO, 2015).

Adotamos split estratificado 70/15/15, seed=42, e padronizacao ajustada somente no treino para evitar vazamento. A rotina de auditoria inclui hash do dataset e registro de versoes. Todos os artefatos sao exportados para validacao independente (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dado o severo desbalanceamento de classes (0,173% positivas), aplicou-se a tecnica de sobreamostragem SMOTE (Synthetic Minority Over-sampling Technique) exclusivamente ao conjunto de treino. Este procedimento visa mitigar o vies do modelo em favor da classe majoritaria, criando instancias sinteticas da classe minoritaria (fraude) para permitir um aprendizado mais robusto dos seus padroes (CHAWLA et al., 2002).

2. Desenvolvimento - Metodos

Comparamos MLP (PyTorch, ReLU, dropout), AE (reconstrucao), LR (baseline linear) e IF (isolamento) com hiperparametros conservadores e foco didatico. Configuracoes estaveis foram priorizadas a tunings agressivos (GOODFELLOW; BENGIO; COURVILLE, 2016; LIU; TING; ZHOU, 2008; KING; ZENG, 2001).

O MLP foi implementado em PyTorch com a seguinte arquitetura: uma camada de entrada com 30 neuronios (correspondente as features PCA), duas camadas ocultas com 16 e 8 neuronios respectivamente, e uma camada de saida com 2 neuronios para classificacao binaria. A funcao de ativacao ReLU foi utilizada nas camadas ocultas, e camadas de Dropout ($p=0.2$) foram inseridas apos cada camada oculta para regularizacao. O modelo foi treinado com o otimizador Adam e a funcao de perda Cross-Entropy, monitorando a metrica F1 na classe positiva.

Avaliamos AUC-ROC/PR, precisao, recall, F1; calibramos thresholds por F1 e F na validacao e aplicamos no teste. Incluimos importancia por permutacao como interpretabilidade pragmatica (SAITO; REHMSMEIER, 2015; DAVIS; GOADRICH, 2006).

Para calibracao de probabilidades, aplicamos Platt Scaling a Regressao Logistica e Temperature Scaling ao MLP, garantindo que as saidas probabilisticas sejam confiaveis e alinhadas com as exigencias de governanca (NICULESCU-MIZIL; CARUANA, 2005; GUO et al., 2017).

Executamos em Google Colab/CPU com PyTorch e scikit-learn, com controle de semente e isolamento de conjuntos para comparacoes justas (GOODFELLOW; BENGIO; COURVILLE, 2016). No protocolo de avaliacao, priorizamos PR na leitura de desempenho; ROC e reportada por tradicao, mas PR captura melhor a utilidade em classe rara (SAITO; REHMSMEIER, 2015; DAVIS; GOADRICH, 2006). Conduzimos tambem um stress test variando a prevalencia positiva (1% 20%) para avaliar sensibilidade a prior shift e estabilidade de F1 (CHEN et al., 2024).

3. Desenvolvimento - Resultados

O MLP obteve o melhor F1 na classe positiva, com a LR ficando muito proxima; AE e IF ficaram abaixo, reforcando a superioridade do supervisionado com rotulos de fraude.

Histogramas mostram separacao de p para a classe 1 no MLP (KING; ZENG, 2001; LIU; TING; ZHOU, 2008; CHEN et al., 2024). A importancia por permutacao destacou V14 e componentes PCA correlatas como determinantes; interpretamos como proxies latentes (BISHOP, 2006; DAL POZZOLO, 2015). Sob prior shift, MLP e LR mantiveram F1 estavel; o IF degradou; o AE apresentou recall razoavel porem precisao baixa (LIU; TING; ZHOU, 2008; CHEN et al., 2024).

Quanto a comparacao de modelos por threshold (validacao/teste), o efeito do limiar e determinante. No threshold padrao de 0,5, o MLP atingiu alta acuracia (98,9% no teste) mas F1 baixo na classe positiva (0,210, com precisao de 0,12 e recall de 0,85), refletindo muitos falsos positivos. Ao elevar o corte para 0,99, o MLP obteve F1 de 0,686 na validacao e 0,747 no teste (precisao 0,674, recall 0,838). O threshold calibrado por $F = 2$ produziu resultados identicos ao corte 0,99 no MLP. A Regressao Logistica no corte 0,99 obteve F1 de 0,659 na validacao e 0,736 no teste proxima ao MLP. O Autoencoder, no corte otimo de F1, alcançou apenas F1 de 0,215 no teste (precisao 0,130, recall 0,608), e o Isolation Forest ficou ainda abaixo, com F1 de 0,179 no teste (precisao 0,111, recall 0,459).

As matrizes de confusao confirmam esse quadro. No teste, o MLP no corte 0,5 produziu 462 falsos positivos e 11 falsos negativos (63 verdadeiros positivos). Ja no corte 0,99, os falsos positivos cairam drasticamente para 30, com 12 falsos negativos e 62 verdadeiros positivos uma reducao de mais de 90% nos alarmes falsos ao custo de apenas um verdadeiro positivo a menos. O Autoencoder no teste apresentou 300 falsos positivos e 29 falsos negativos (45 verdadeiros positivos), evidenciando sua menor precisao.

O stress test com variacao de prevalencia (prior shift, de 1% a 20% de positivos) confirmou a estabilidade dos modelos supervisionados. O MLP manteve F1 entre 0,71 e 0,75 ao longo de todos os cenarios de prevalencia (0,747 em 1%, 0,744 em 5%, 0,713 em 10% e 0,734 em 20%). A LR exibiu comportamento semelhante, oscilando entre 0,68 e 0,78 (chegando a superar o MLP em 10%, com F1 de 0,780). O Isolation Forest permaneceu consistentemente baixo (F1 entre 0,179 e 0,194), demonstrando degradacao e inadequacao ao problema.

A importancia por permutacao, calculada na validacao, identificou a variavel V14 como de longe a mais determinante, com importancia de 0,073 cerca de cinco vezes superior a segunda colocada, V2 (0,015), seguida por V22 (0,014), V24 (0,007) e V25 (0,006).

Diversas variaveis apresentaram importancia proxima de zero ou levemente negativa (por exemplo V13, com -0,004, e V26, com -0,002), indicando contribuicao marginal ou nula para a discriminacao da classe positiva.

Quanto as figuras principais, observa-se que a curva de Precisao-Recall (PR) do MLP domina as demais abordagens em praticamente todo o espaco de recall. Esse resultado e particularmente relevante em cenarios de forte desbalanceamento, como no presente estudo, em que a AUC-PR fornece uma avaliacao mais confiavel da capacidade do modelo de identificar fraudes. O desempenho superior do MLP nesta metrica confirma sua adequacao para o problema, evidenciando maior robustez na deteccao da classe minoritaria em comparacao com os demais modelos.

4. Discussao

A performance robusta do MLP, que nao apenas superou os modelos nao supervisionados, mas tambem competiu de perto com a Regressao Logistica, demonstra sua capacidade de aprender padroes nao-lineares complexos. O uso estrategico de Dropout foi fundamental para controlar o risco de overfitting, garantindo que o modelo generalizasse bem para o conjunto de teste, como evidenciado pela proximidade entre as metricas de validacao e teste.

Em dados tabulares com PCA, a LR ja captura muito do sinal; o MLP adiciona nao-linearidade util aumentando F1 e AUC-PR (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; GOODFELLOW; BENGIO; COURVILLE, 2016).

A escolha de threshold deve refletir custo de negocio; recomenda-se manter cortes $F1$ -otimo e F , revisados periodicamente conforme drift (SAITO; REHMSMEIER, 2015; DAVIS; GOADRICH, 2006).

Para governanca, recomenda-se calibracao de probabilidades e validacao de estabilidade de importancia (bootstrap) e SHAP para explicacoes locais (NICULESCU-MIZIL; CARUANA, 2005; GUO et al., 2017; BISHOP, 2006).

Limitamo-nos a um dataset; a generalizacao requer validacao externa e possivel ajuste fino de hiperparametros e limiares (CHEN et al., 2024). Nao realizamos busca exaustiva nem calibracao nesta versao; priorizamos replicabilidade e clareza didatica conforme diretrizes do curso. Adicionalmente, os dados PCA podem introduzir vieses latentes nao explorados; sugerimos experimentos futuros com undersampling ou ensembles (como Random Forest com SMOTE) para maior robustez.

Apesar dos resultados promissores, algumas limitacoes devem ser consideradas para aplicacao pratica. Em ambientes reais de transacoes financeiras, o custo computacional do treinamento de redes neurais profundas pode ser elevado, especialmente quando ha necessidade de reprocessamento frequente de grandes volumes de dados. Alem disso, a ocorrencia de concept drift isto e, a mudanca gradual no padrao das transacoes legitimas e fraudulentas ao longo do tempo pode comprometer a generalizacao do modelo, exigindo monitoramento continuo e atualizacoes periodicas para manter sua acuracia. Dessa forma, embora o MLP tenha demonstrado resultados superiores, sua adocao em producao deve ser acompanhada de estrategias de manutencao e validacao continua.

5. Consideracoes Finais

Recomendamos o MLP como modelo principal e a LR como baseline explicavel; sugere-se threshold sensivel a custo, calibracao e monitoramento de drift em producao (KING; ZENG, 2001; NICULESCU-MIZIL; CARUANA, 2005; GUO et al., 2017). Para aplicacoes reais, integre o modelo a sistemas de monitoramento em tempo real, com alertas baseados em thresholds calibrados para minimizar impactos operacionais. Disponibilizamos notebook unico, figuras/tabelas exportadas e sumario JSON (versoes/seed/hash) garantindo auditoria ponta-a-ponta (FLORES, 2025).

6. Referencias

- SAITO, T.; REHMSMEIER, M. The Precision-Recall Plot Is More Informative than the ROC Plot on Imbalanced Datasets. PLoS ONE, 2015.
- DAVIS, J.; GOADRICH, M. The Relationship Between Precision-Recall and ROC Curves. In: ICML, 2006.
- KING, G.; ZENG, L. Logistic Regression in Rare Events Data. Political Analysis, 2001.
- NICULESCU-MIZIL, A.; CARUANA, R. Predicting Good Probabilities with Supervised Learning. In: ICML, 2005.
- GUO, C.; PLEISS, G.; SUN, Y.; WEINBERGER, K. On Calibration of Modern Neural Networks. In: ICML, 2017.
- LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation Forest. In: ICDM, 2008.
- CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 2002.
- DAL POZZOLO, A. Adaptive Machine Learning for Credit Card Fraud Detection. PhD Thesis, 2015.
- ULB/WORLDDLINE. Credit Card Fraud Dataset. Kaggle mirror, 2013.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, 2016.
- BISHOP, C. Pattern Recognition and Machine Learning. Springer, 2006.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. Springer, 2009.

MURPHY, K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

CHEN, W. et al. A survey on imbalanced learning: latest research. Artificial Intelligence Review, 2024.

HERNANDEZ AROS, L. et al. Financial fraud detection through ML. Humanities and Social Sciences Communications, 2024.

CHERIF, A. et al. Credit card fraud detection in the era of disruptive technologies. JISA, 2023.

IBM Skills Network. AI Development Estudo de Caso (Diretrizes). 2025.

IBM Skills Network. Unit 3.x Laboratorios de Metricas, Treinamento e Avaliacao. 2025.

FLORES, Carlos Ulisses. Notebook

estudo_caso_fraude_cartao_pytorch_v3p2_final_full.ipynb. Colab, 2025. Acesso em: 16/08/2025.

Canonical URL: <https://ulissesflores.com/research/2025-fraud-detection-mlp>

Primary PDF URL: <https://ulissesflores.com/deep-research/2025-fraud-detection-mlp/deep-research.pdf>

Legacy PDF URL: <https://ulissesflores.com/research/2025-fraud-detection-mlp.pdf>

Generated from UPKF at 2026-02-21