Construção de um Modelo Orientado a Dados para Detecção de Fraudes em Cartões de Crédito utilizando Dados Sintéticos

Alexandre C. B. dos Santos[1], Roger de S. Passos[2], Luis Domingues T. J. Tarrataca[2], Douglas de O. Cardoso[3], Diego B. Haddad[2], Felipe da R. Henriques[1,2]

[1]PPCIC, Cefet/RJ - Rio de Janeiro - RJ - Brasil

[2]Cefet/RJ - Petrópolis - RJ - Brasil

[3]Center of Linguistics, University of Porto - Porto - Portugal

{alexandre.santos,roger.passos}@aluno.cefet-rj.br,

{luis.tarrataca,diego.haddad,felipe.henriques}@cefet-rj.br,

docardoso@letras.up.pt

Abstract. *Credit card transaction fraud is a global challenge, resulting in significant financial losses. This work proposes a synthetic data simulator for transactions to replicate the dynamics of real-world data. These data were used to create models based on classification algorithms and anomaly detection, capable of identifying fraudulent transactions. Challenges such as sequential modeling, context change, delayed feedback, and data peculiarities were addressed. The Random Forest algorithm stood out, detecting 76.7% of frauds with a precision of 96.4%.*

Resumo. *Fraudes em transações com cartões de crédito são um desafio global, resultando em grandes prejuízos financeiros. Este trabalho propõe um simulador de dados sintéticos de transações para replicar a dinâmica de dados reais. Esses dados foram usados para criar modelos baseados em algoritmos de classificação e detecção de anomalias, capazes de identificar fraudes. Desafios como modelagem sequencial, mudança de contexto, feedback atrasado e peculiaridades dos dados foram abordados. O algoritmo Random Forest destacou-se, detectando 76,7% das fraudes com 96,4% de precisão.*

1. Introdução

Fraudes em transações com cartões de crédito representam um desafio significativo para instituições financeiras, empresas e indivíduos, com perdas globais que atingiram 32,34 bilhões de dólares em 2021, segundo a Nilson [nil 2022]. No Brasil, a Serasa [ser 2022] destaca que mais da metade das tentativas de golpes em maio de 2022 ocorreram no segmento de bancos e cartões, e uma reportagem do G1 [g1 2023] de 2023 revela que o país é o quinto mais afetado por roubo e venda de cartões na *dark web*.

Dada a gravidade do problema, há um esforço conjunto de empresas e pesquisadores para aprimorar a detecção de fraudes, apesar dos desafios impostos pela complexidade do problema e pela escassez de dados públicos. Este trabalho busca enfrentar esses desafios ao sistematizar a construção de modelos de detecção de fraudes e ao propor melhorias no simulador de dados sintéticos de transações desenvolvido por [Le Borgne et al. 2022].

Os objetivos incluem aprimorar o simulador para melhor refletir a complexidade dos dados reais, desenvolver modelos robustos de aprendizado de máquina e comparar os resultados obtidos em diferentes abordagens de classificação e detecção de anomalias.

2. Simulador de Dados Sintéticos

O simulador de dados sintéticos utilizado neste trabalho é uma evolução direta daquele disponível em [Le Borgne et al. 2022]. Os autores deste último afirmam que seu simulador possui um *design* simples comparado à dinâmica complexa de dados reais de pagamentos com cartões. Neste trabalho, o simulador foi aprimorado com três novas características: diferenciação do tipo de transação, um novo cenário de fraude e melhorias no esquema de localizações.

É preciso notar que há uma diferença clara entre transações de cartões de crédito feitas presencialmente (CP - Cartão Presente) e *online* (CNP - Cartão não Presente). Os clientes realizam, em geral, transações CP em um número limitado de comerciantes sujeitos a um horário de funcionamento. Além disso, transações CP são consideradas mais seguras devido à presença física do cliente e ao uso de cartões com tecnologia EMV (*Europay, Mastercard, and Visa*). Por outro lado, as transações CNP podem ser realizadas sem restrições de local ou horário e representam a vasta maioria das fraudes [ban 2023]. Tais distinções foram, então, acrescentadas ao simulador com a criação do atributo do tipo da transação.

Um novo cenário de fraude também foi acrescentado. Assim, o simulador passou a contar com o principal tipo de fraude CP [ban 2023]: fraude por cartão perdido ou furtado. Esse cenário visa representar o comportamento de um fraudador que gasta o máximo possível o mais rápido possível, antes que o cartão seja bloqueado. Para identificar esse tipo de fraude, seria pertinente a criação de atributos capazes de registrar tanto a quantidade quanto o valor esperado das transações do cliente em um curto período de tempo.

Por fim, implementou-se um esquema mais realista de localizações. No simulador inicial, as localizações eram escolhidas uniformemente a partir de um *grid* 100×100. Agora, as coordenadas geográficas do Brasil são utilizadas, tomando como principal fonte as informações de um repositório de dados sobre os municípios brasileiros[1]. A Figura 1 exibe a abrangência territorial das localizações geradas.

Nesse novo procedimento, primeiro escolhe-se um município de forma aleatória, ponderado pelo tamanho da população. Quanto mais populoso o município, maior a probabilidade de ser escolhido. Determinado o município, uma localização específica é então definida dentro de um raio de 12 km de seu centro geográfico. Este valor para o raio foi estabelecido como uma aproximação grosseira da área mediana dos municípios brasileiros[2].

A fim de abranger as três principais localizações relacionadas à transações com cartões de crédito, a localização de entrega do produto foi adicionada. Sua definição parte do tipo da transação. Em transações CP, o cliente está fisicamente presente para tomar

[1] `https://github.com/mapaslivres/municipios-br` (acesso em 01/06/2024).

[2] `https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html` (acesso em 01/06/2024).

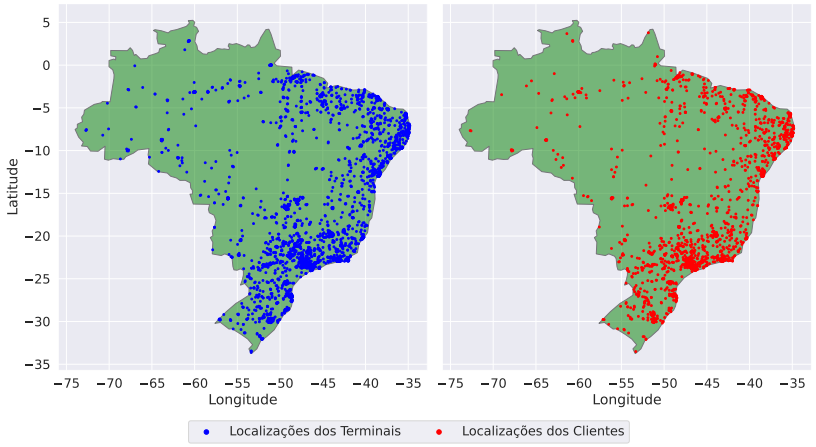

Figura 1. Amostra de 20% das localizações geradas pelo simulador.

posse do item adquirido, então é mais raro que a localização de entrega seja diferente da localização do terminal de compra. Já em transações CNP, em que a localização de entrega é vital, presume-se que, na maioria dos casos, coincidirá com o endereço de cobrança do cliente.

Aproveitando-se disso, uma estratégia para evitar fraudes seria bloquear transações que fogem do padrão de localização de entrega, utilizando, por exemplo, o *Address Verification Service* [Wong et al. 2012]. Isso, no entanto, acarretaria em perdas financeiras, uma vez que clientes legítimos podem optar por enviar suas compras para outras localizações. Essa variabilidade foi incorporada à escolha das localizações de entrega no simulador e deve ser considerada ao criar atributos para a detecção das fraudes.

Com essas contribuições, almejou-se aproximar mais os dados gerados pelo simulador de um *dataset* real de transações. Essas melhorias foram concebidas com o propósito adicional de preservar a interpretabilidade, assegurando que o processo de geração das transações e fraudes permanecesse transparente. Cabe ressaltar, por último, que todos os parâmetros empregados na construção do *dataset* foram escolhidos empiricamente. Assim, ao mesmo tempo que há flexibilidade para ajustá-los conforme necessário, é preciso notar que o *dataset* final, sendo sensível a esses valores, pode ser substancialmente distinto.

3. Modelo Orientado a Dados

Com o conjunto de dados de transações obtidos a partir do simulador descrito na seção anterior, pode-se prosseguir para a metodologia de desenvolvimento de um modelo orientado a dados para a detecção das fraudes. Dois tipos de algoritmos de aprendizado de máquina foram utilizados: classificadores e detectores de anomalias. Entre os classificadores, foram empregados o *Random Forest*, *K-Nearest Neighbors* e *Logistic Regression*. Para a detecção de anomalias, foram utilizados o *Isolation Forest* e *Elliptic Envelope*.

Na etapa de engenharia de atributos, os dados foram transformados preservando informações críticas, como o valor numérico das transações, a indicação de fraude e o tipo de transação (CP ou CNP). Distâncias geográficas entre as localizações de cobrança e entrega foram calculadas, e técnicas de agregação foram utilizadas para capturar hábitos de consumo dos clientes, considerando o total e a média das transações em diferentes

períodos (1, 7 e 30 dias), segmentados pelo tipo de transação. Padrões de distâncias entre as localizações de cobrança e entrega foram capturados por cliente e tipo de transação, e o risco associado a terminais foi calculado com base na proporção de fraudes, considerando um atraso de 7 dias para a confirmação das transações.

Devido ao *feedback* atrasado para obtenção do histórico de transações rotulado, os dados imediatamente anteriores ao período de teste não podem ser utilizados para treinamento e, portanto, precisam ser removidos. Além disso, dados mais antigos também são removidos em função da mudança de contexto, assegurando que o conjunto de treinamento seja composto das mais recentes transações. Ao considerar todos esses aspectos, a divisão de dados da Figura 2 foi empregada.

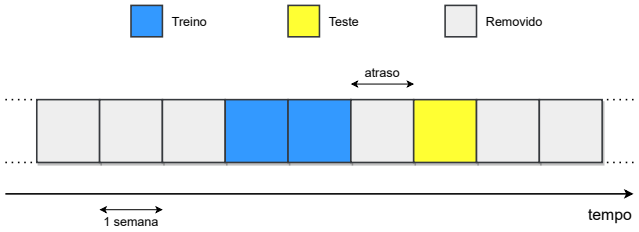

Figura 2. Divisão dos dados considerando o aspecto temporal das transações. Utilizou-se 14 dias de dados de treinamento e 7 dias de dados de teste, com um atraso de 7 dias entre esses conjuntos.

Técnicas padrão de validação cruzada não são adequadas para dados com uma componente temporal, pois desconsideram a ordem dos dados. Em vez disso, optou-se por utiliza a abordagem *prequential* [Gama et al. 2014, Cerqueira et al. 2020, Le Borgne et al. 2022], onde, na primeira iteração, os dados são divididos conforme ilustrado na Figura 2 e obtém-se uma avaliação do modelo. Em iterações subsequentes, os conjuntos de treinamento e teste são deslocados, permitindo que dados mais antigos sejam substituídos por amostras mais recentes para teste. Esse processo é repetido n vezes, e o resultado final é a média dos resultados de cada iteração.

A abordagem *prequential* foi adotada tanto na etapa de treino e validação dos modelos de aprendizado de máquina quanto na fase de treino e teste, que consistiu na avaliação e comparação entre os modelos definitivos.

Por fim, realizou-se a construção dos modelos de aprendizado de máquina, que consistiu em um procedimento composto por duas etapas aninhadas. Para cada algoritmo, foram exploradas diversas combinações de estratégias de pré-processamento, tais como, normalização *min-max*, padronização, análise de componentes principais e reamostragem. Além disso, foram considerados subconjuntos dos principais hiperparâmetros desses algoritmos, para serem utilizados durante a otimização de hiperparâmetros por meio da técnica de *grid search*.

Ao iterar por diferentes estratégias de pré-processamento e configurações de hiperparâmetros, buscou-se obter o modelo mais eficaz para cada algoritmo priorizando o desempenho no conjunto de validação, com ênfase na métrica *Average Precision*. Durante esse processo, todos os modelos foram treinados utilizando a técnica de aprendizado em lotes. As implementações foram realizadas com o auxílio da biblioteca *scikit-learn* [Pedregosa et al. 2011] em sua versão 1.3.2.

4. Avaliação Experimental

4.1. Métricas de Desempenho

Para determinar a classe de saída dos algoritmos de aprendizado de máquina, foi utilizado um limiar de decisão fixo, onde fraudes são consideradas positivas e transações genuínas, negativas. A matriz de confusão quantifica erros e acertos dos modelos, permitindo derivar métricas como Taxa de Falsos Positivos (FPR), Taxa de Verdadeiros Positivos (TPR), Precisão, *G-Mean* e *F1-Score*. Também foram empregadas curvas ROC e *Precision-Recall*, e suas áreas (AUC ROC e *Average Precision* - AP) para avaliar a eficácia dos modelos em diferentes limiares de decisão.

4.2. Avaliação dos Modelos de Aprendizado de Máquina

As Figuras 3 e 4 mostram, respectivamente, as curvas ROC e *Precision-Recall* para cada modelo, representando uma síntese das iterações da divisão *prequential*. A curva ROC foi interpolada entre resultados intermediários, enquanto para a *Precision-Recall*, as pontuações das iterações foram concatenadas, uma vez que a interpolação seria inadequada neste gráfico [Davis and Goadrich 2006]. Para uma análise comparativa mais detalhada, as métricas AUC ROC e AP foram consideradas. A Figura 5 revela que os valores de AUC ROC são similares entre os modelos, mas a média da AP permite ordenar o desempenho, destacando o classificador *Random Forest*.

Para uma análise mais tangível do desempenho dos modelos, foram consideradas métricas que dependem de um limiar, sendo que a definição deste pode variar conforme os objetivos. Neste trabalho, optou-se por maximizar a métrica *F1-Score*, que busca o equilíbrio entre precisão e sensibilidade na detecção de fraudes. A escolha do limiar foi feita para alcançar o maior valor possível de *F1-Score*, o que se reflete no ponto mais próximo de $(1, 1)$ nas curvas da Figura 4.

Os resultados apresentados na Tabela 1 mostram que todos os três classificadores atingem excelentes valores de precisão quando o critério de *F1-Score* é utilizado. Por exemplo, o classificador *Random Forest* consegue identificar aproximadamente 76,7% das fraudes com uma precisão notável de cerca de 96,4%, o que resulta em uma taxa muito baixa de falsos positivos.

Tabela 1. Resultados obtidos para o limiar que maximiza a métrica *F1-Score*.

Algoritmo	Métricas				
	FPR	TPR	Precisão	G-Mean	F1-Score
Random Forest	$\mathbf{0,0 \pm 0,0}$	$\mathbf{0,767 \pm 0,017}$	$\mathbf{0,964 \pm 0,013}$	$\mathbf{0,876 \pm 0,01}$	$\mathbf{0,854 \pm 0,014}$
Logistic Regression	0,001 ± 0,0	0,685 ± 0,031	0,864 ± 0,043	0,827 ± 0,019	0,763 ± 0,012
K-Nearest Neighbors	$\mathbf{0,0 \pm 0,0}$	0,685 ± 0,02	0,9 ± 0,024	0,828 ± 0,012	0,778 ± 0,012
Isolation Forest	0,002 ± 0,001	0,425 ± 0,033	0,474 ± 0,042	0,65 ± 0,025	0,446 ± 0,018
Elliptic Envelope	0,002 ± 0,0	0,413 ± 0,038	0,587 ± 0,019	0,641 ± 0,03	0,484 ± 0,028

5. Conclusões

A detecção de fraudes em cartões de crédito é um desafio complexo que envolve múltiplas etapas. Um simulador de transações foi desenvolvido para gerar dados realistas, seguido

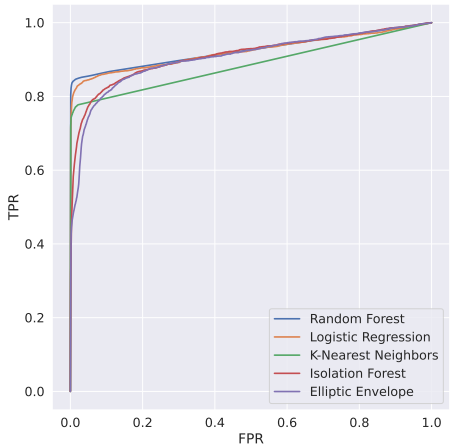

Figura 3. Curva ROC dos modelos, sumarizada para cada iteração da divisão *Prequential*.

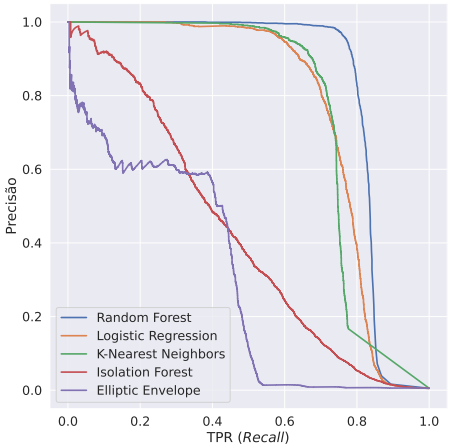

Figura 4. Curva *Precision-Recall* dos modelos, sumarizada para cada iteração da divisão *Prequential*.

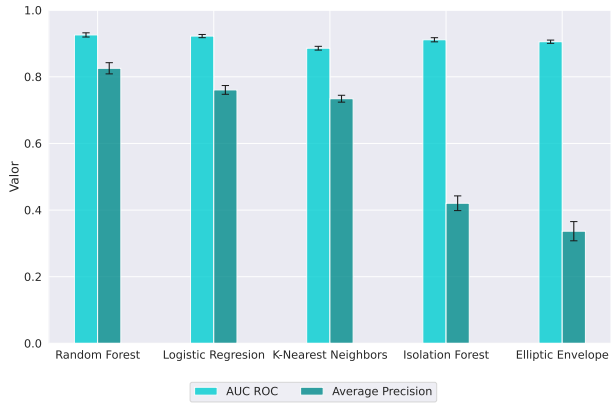

Figura 5. Média e desvio padrão das métricas AUC ROC e AP para cada modelo.

pela aplicação de técnicas de engenharia de atributos para criar um conjunto de dados mais informativo. Esses dados foram divididos e utilizados na construção e avaliação de modelos de classificação e detecção de anomalias, destacando-se o *Random Forest*, que detectou 76,7% das fraudes com 96,4% de precisão. No entanto, os detectores de anomalias tiveram desempenho inferior aos classificadores, possivelmente devido à menor quantidade de informações disponíveis para o aprendizado.

Para melhorar os detectores de anomalias, uma abordagem potencial seria calcular as pontuações de anomalia dentro de grupos de transações ou clientes após uma *clusterização*, permitindo uma identificação mais específica em diferentes contextos. Além disso, o aprimoramento do simulador de dados pode ser alcançado com um ajuste fino dos parâmetros e a especificação do padrão temporal de compra de cada cliente e dos cenários de fraude. A incorporação de um grau mais elevado de *drifts* nesses padrões pode tornar a dinâmica dos dados ainda mais realista.

Referências

(2022). Mais de 50% das tentativas de fraude são no segmento de bancos e cartões, aponta serasa experian. Disponível em: https://www.serasaexperian.com.br/sala-de-imprensa/analise-de-dados/mais-de-50-das-tentativas-de-fraude-sao-no-segmento-de-bancos-e-cartoes-aponta-serasa-experian/. Acesso em: 01/06/2024.

(2022). Nilson report. Disponível em: https://nilsonreport.com/. Acesso em: 01/06/2024.

(2023). Mais de 140 mil cartões foram roubados no brasil e vendidos na 'dark web' em 2023, diz pesquisa. Disponível em: https://g1.globo.com/economia/noticia/2023/05/29/mais-de-140-mil-cartoes-foram-roubados-no-brasil-e-vendidos-na-dark-web-em-2023-diz-pesquisa.ghtml. Acesso em: 01/06/2024.

(2023). Report on card fraud in 2020 and 2021. Disponível em: https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202305~5d832d6515.en.html. Acesso em: 01/06/2024.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109:1997–2028.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Cohen, W. W. and Moore, A. W., editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 233–240. ACM.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37.

Le Borgne, Y.-A., Siblini, W., Lebichot, B., and Bontempi, G. (2022). *Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*. Université Libre de Bruxelles.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Wong, N., Ray, P., Stephens, G., and Lewis, L. (2012). Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results. *Inf. Syst. J.*, 22(1):53–76.