II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Detecção de fraude no comércio eletrônico brasileiro

Rafael Belmiro Cristovao[1], Gustavo Carlos Buscaglia[2]
ICMC-USP

1 Introdução

O e-commerce brasileiro segue em rápida expansão e faturou mais de R$ 161 bilhões em 2021 [5]. O cartão de crédito é uma das principais formas de pagamento utilizadas na modalidade, entretanto o crescimento de sua popularidade o tornou visado pelos fraudadores. A fraude é uma preocupação constante dos e-commerces brasileiros, que sofre mais de R$ 3.6 mil em tentativas de fraudes por minuto [1]. Por este motivo, o tema chama atenção de pesquisadores e trabalhos científicos e se tornou um tópico de pesquisa frequente com o crescimento do interesse em big data e aprendizado de máquina [6]. Existem diversos desafios encontrados na literatura para a criação de modelos de aprendizado de máquina na previsão de fraudes, alguns exemplos são o desbalanceamento entre as classes, a tendência não estacionária da distribuição e a falta de dados públicos para análise. O desbalanceamento entre classes acontece pois o número de compras não fraudulentas é muito maior do que de tentativas de fraudes, já que a quantidade de bons consumidores é muito grande. Já a distribuição não estacionária se dá pela mudança de comportamento de bons consumidores, por exemplo devido à períodos promocionais, e de fraudadores, dado que eles podem mudam seu comportamento a fim de burlar os sistemas de detecção [8], esse problema também é conhecido como *concept drift*. Por fim, os dados para pesquisas acadêmicas sobre o tema são escassos devido à características sensíveis das informações, pois existem muitas informações pessoais, como CPF e e-mail, e informações sensíveis, como informações do cartão de crédito utilizado na compra. O presente trabalho tem como objetivo comparar diferentes algoritmos de aprendizado de máquina em uma base real de compras online e estudar o impacto do *concept drift* na predição.

2 Metodologia

Utilizaremos uma base de dados real de compras online de uma loja de e-commerce composta por 11.211.709 transações realizadas entre julho/2021 e outubro/2021, sendo 419.895 com marcações de fraude, 3.745% do total de transações. As marcações podem ocorrer pelo processo

[1] rafael.cristovao@usp.br
[2] gustavo.buscaglia@icmc.usp.br

de revisão manual de transações selecionadas ou pela notificação do dono do cartão de crédito, processo conhecido como *chargeback*. No primeiro caso, a resposta de fraude é conhecida, no geral, em até 48 horas da data da transação [2], já o processo de *chargeback* é mais demorado e pode levar meses.

Além da marcação de fraude, a base de dados possui 152 variáveis que serão utilizadas para treinamento dos modelos. Entre elas estão variáveis categóricas, como domínio do e-mail utilizado na conta, CEP de entrega, etc e variáveis numéricas, como valor da compra, valor do frete, quantidade de compras realizadas nos últimos 7 dias, etc.

Pela natureza estocástica do padrão de fraude, para comparação dos algoritmos a base de dados total foi dividida em cinco lotes temporais de acordo com a figura 1, onde os lotes $Lote_i$ $i \in (1, 2, 3, 4)$ foram usados como períodos de treinamento para os modelos, que foram testados no lote imediatamente posterior, totalizando 4 treinos para cada algoritmo.

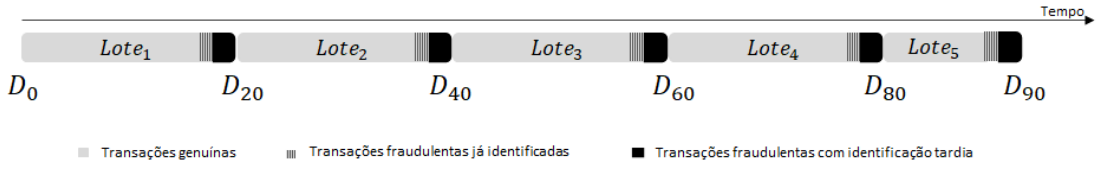

Figura 1: Divisão da base de dados em lotes

Além disso, para o período de treinamento retiramos as notificações de fraudes que não tinham sido notificadas até a data de treino, evitando informação futura no treinamento. Para avaliação no período de teste, todas as notificações foram consideradas.

Antes do treinamento dos algoritmos, na etapa de pré-processamento de cada lote, os seguintes passos foram feitos:

- Remoção de variáveis com mais que 40% de valores nulos
- Atribuição de categoria específica para os valores nulos das variáveis categóricas
- Atribuição do valor médio para os valores nulos das variáveis numéricas
- Transformação das variáveis categóricas nos pesos de evidência de cada categoria
- Normalização das variáveis numéricas

Por fim, utilizamos a biblioteca *sklearn* da linguagem de programação *python* para o treinamento dos algoritmos Máquina de vetores de suporte (SVM), Regressão Logística (RL), Rede Neural (RN) e Florestas Aleatórias (RF) e o *framework LightGBM* [4] para o treinamento do modelo *Gradient Boosting Decision Tree* (LGBM). Para encontrar os melhores hiperparâmetros de cada algoritmo, utilizamos o método *GridSearchCV* da biblioteca *sklearn*, que implementa uma busca em grade nos espaço de parâmetros de interesse.

3 Resultados e Discussões

Escolher uma boa medida para avaliação de modelos de prevenção à fraude muitas vezes não é uma tarefa trivial. O desbalanceamento das classes de interesse torna algumas medidas clássicas de avaliação de modelos de classificação, como acurácia, pouco úteis [2]. Nesse caso, a área sob a curva ROC é uma medida bem aceita pois considera as distribuições das duas classes de interesse para todos possíveis pontos de corte. Outra medida amplamente utilizada é a estatística KS, derivada do teste de hipótese não paramétrico de Kolmogorov-Smirnov [3] [7], que compara as distribuições de probabilidade acumuladas das transações fraudulentas e não fraudulentas. No cenário decrito na figura 1, o modelo LGBM teve melhor resultado para todos os lotes, seguido da RF.

Tabela 1: Métricas do período de teste para cada lote

	AucRoc					KS				
Lote	LGBM	LR	RF	RN	SVM	LGBM	LR	RF	RN	SVM
1	0.979	0.967	0.977	0.975	0.954	0.852	0.816	0.844	0.836	0.779
2	0.974	0.954	0.970	0.967	0.938	0.829	0.769	0.814	0.804	0.729
3	0.978	0.967	0.977	0.970	0.955	0.850	0.815	0.843	0.827	0.787
4	0.977	0.962	0.973	0.956	0.943	0.844	0.792	0.834	0.783	0.748

A RF teve melhor performance no conjunto de treinamento para ambas as métricas, mas perdeu performance na generalização para o período de teste, já o modelo LGBM teve uma performance um pouco pior do que a RF no conjunto de treinamento mas conseguiu melhor generalização, alcançando as melhores médias no conjunto de dados de teste. O SVM foi o algoritmo que teve pior desempenho, pelo tempo de processamento só foi testado o núcleo linear, outros núcleos podem melhor a performance do classificador, mas exigem um algo tempo computacional para convergência. A média das métricas para os conjuntos de treinamentos e testes estão na tabela 2.

Tabela 2: Média das métricas

Modelo	Treino		Teste	
	AucRoc	KS	AucRoc	KS
LGBM	0.9853	0.8762	**0.9770**	**0.8438**
RF	**0.9898**	**0.8987**	0.9743	0.8337
RN	0.9765	0.8399	0.9670	0.8124
RL	0.9730	0.8274	0.9625	0.7978
SVM	0.9608	0.7941	0.9475	0.7606

4 Conclusão e Próximos Passos

A dinâmica das tentativas de fraudes de cartão de crédito é um ambiente bastante complexo devido a quantidade de fatores que podem impactar sua avaliação. Nesse contexto, podemos observar que os métodos baseados em árvores de decisão obtiveram melhores resultados para o conjunto de transações avaliadas, sendo o LGBM o algoritmo com maior performance. Esses métodos conseguiram captar melhor a relação não linear do evento no conjunto de variáveis disponíveis, já o SVM utilizando núcleo linear foi o pior em representar a tendência não linear do evento e tiveram pior desempenho.

Como próximos passos, estudaremos o comportamento do modelo que teve melhor desempenho, o LGBM, ao longo do tempo a fim de avaliar a presença de *concept drift* na base de dados e o impacto do retreino na performance da solução.

Referências

[1] Clearsale. E-commerce brasileiro sofre mais de R$ 3,6 mil em tentativas de fraude por minuto. Disponível em: https://www.ecommercebrasil.com.br/noticias/e-commerce-brasileiro-sofre-mais-de-r-36-mil-em-tentativas-de-fraude-por-minuto/. Acesso em: 21 jul. 2022.

[2] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G. *Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797.

[3] Fidel Beraldi. Atualização dinâmica de modelo de regressão logística binária para detecção de fraudes em transações eletrônicas com cartão de crédito. Dissertação de Mestrado em Ciências, Universidade de São Paulo (USP), 2014.

[4] Microsoft Corporation. *LightGBM Release 3.3.2*. Disponível em: https://lightgbm.readthedocs.io/_/downloads/en/v3.3.2/pdf/. Acesso em: 21 jul. 2022.

[5] Neotrust. E-commerce brasileiro cresce 27% e fatura R$ 161 bilhões em 2022. Disponível em: https://www.ecommercebrasil.com.br/noticias/neotrust-e-commerce-fatura-2021/. Acesso em: 21 jul. 2022.

[6] Niu, X., Wang, L., Yang, X. *A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised*, 2019

[7] Oliveira, P. H. M. A. etecção de fraudes em cartões: um classificador baseado em regras de associação e regressão logística. Dissertação de Mestrado em Ciências, Universidade de São Paulo (USP), 2016.

[8] Soemers, D. J. N. J., Brys, T., Driessens, K., Winands, M. H. M., Nowé, A. *Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees. 32nd AAAI Conference on Artificial Intelligence*, AAAI 2018, 1, 7831–7836.