Comparativo de técnicas de inteligência artificial explicável na detecção de fraudes em transações com cartão de crédito

Gabriel Mendes de Lima[1],
Paulo Henrique Pisani[1]

Universidade Federal do ABC (UFABC)[1]

Introdução

- Modelos de aprendizado de máquina utilizados na avaliação da legitimidade de transações com cartão de crédito **[Makki et al. 2019, Chaudhary et al. 2012]**;

- Número elevado de transações, inviável de lidar apenas com analistas humanos.

Introdução

Modelos de aprendizado automática podem ser:

- **Naturalmente interpretáveis**
- **Caixa-preta**

Introdução

- Sistemas não auditáveis, não são adequados a setores altamente regulados **[Bussmann et al. 2021]**;
- A União Europeia e seus países membros exigem isso por lei **[European Union 2016]**;
- Segurança por meio da interpretabilidade **[Doshi-Velez and Kim 2017]**.

Introdução

Contexto estudado

- Construir modelos que:
  - Com base em um conjunto de transações, consigam identificar transações fraudulentas das legítimas;

Introdução

Objetivo

- Com base nos modelos criados:
  - Comparar técnicas de inteligência artificial explicável e os resultados das mesmas;
  - LIME **[Ribeiro et al. 2016]**;
  - SHAP **[Lundberg and Lee 2017]**.

Introdução

Desafios

- Quando podemos dizer que o modelo é interpretável ou que uma explicação é boa o suficiente? Não há consenso sobre isso **[DOSHI-VELEZ; KIM, 2017]**.

Introdução

Desafios

- O desbalanceamento de classe acontece quando há mais observações de uma classe do que de outra **[MAKKI et al., 2019]**.

Trabalhos Relacionados

- **Ji, Helldin e Steinhauer (2021)** investigaram por meio de uma pesquisa quantitativa a eficiência da aplicação dos algoritmos LIME e SHAP;

- **Psychoula et al. (2021)** investigou a aplicação dos algoritmos LIME e SHAP para gerar explicações no problema de detecção de fraudes com cartões de crédito em tempo real;

- **Hsin et al. (2021)** argumentou que abordagens tradicionais baseadas em regras para detecção de fraudes com cartões de crédito não são efetivas para este problema;

Trabalhos Relacionados

- **Wu and Wang. (2021)** propuseram um framework para detecção de fraudes com cartões de crédito

- **Chaquet-Ulldemolins et al. (2022)** apresentou uma proposta de metodologia interpretável e agnóstica de modelo para detecção de fraudes com cartões de crédito utilizando autoencoders.

Metodologia

Comparando técnicas de IA Explicável

- **Avaliação por meio de função**
- Avaliação humana
- Avaliação por meio de aplicação

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. 2017

Metodologia

Conjuntos de dados

- Kaggle fraud detection
  - 284.807 transações, 492 fraudes;
  - 30 preditores;
  - 28 transformados com PCA, Time e Amount não foram modificados;
  - 0,017% de exemplos da classe positiva.

Metodologia

Conjuntos de dados

- IEEE-CIS fraud detection
  - Cerca de 500.000 transações;
  - 433 preditores, só foram usados 19;
  - 3,50% de transações fraudulentas.

Metodologia

Conjuntos de dados

- Credit card transaction
  - Cerca de 24 milhões de registros, apenas 480.000 utilizados;
  - 12 preditores;
  - 0,012% de transações fraudulentas.

Metodologia

Configuração dos experimentos

1. Pré-processamento;
2. Separação dos dados;
3. Classificação;
4. Obtenção de métricas;
5. Aplicação dos explicadores.

Metodologia

Pré-processamento

- Remoção de colunas não utilizadas;
- Colunas foram renomeadas;
- Imputation **[Moepya et al. 2016]**;
- Transformação dos dados categóricos **[Bourdonnaye and Daniel 2021]**;

[1]MOEPYA, S. O.; AKHOURY, S. S.; NELWAMONDO, F. V.; TWALA, B. The Role Of Imputation In Detecting Fraudulent Financial Reporting. 2016. 333-356 p.
[2]BOURDONNAYE, F. D. L.; DANIEL, F. Evaluating categorical encoding methods on a real credit card fraud detection database. 2021.

Metodologia

Separação dos dados

- Validação cruzada estratificada
  - 5 folds, 1 repetição;
  - RepeatedStratifiedKFold(n_splits=5, n_repeats=1, random_state=1)[1] ;
  - Aplicação dos métodos de imputation;

[1]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html

Metodologia

Separação dos dados

- Transformar dados categóricos em numéricos;
- Aplicação do método SMOTE **[Alfaiz and Fati 2022]**;
- Criação dos explicadores.

Metodologia

Algoritmos de classificação

- Árvores de decisão **(MAKKI et al., 2019)**.
- Regressão logística **(CHAUDHARY; YADAV; MALLICK, 2012)**.
- Floresta aleatória **(POZZOLO et al., 2018)**.
- Support vector machines (SVMs) **(MAKKI et al., 2019)**.
- Redes neurais artificiais **(ALFAIZ; FATI, 2022)**.

Metodologia

Aplicação dos explicadores

- **SHAP:** Explainer(model, X_train_smote, seed=1)[1]
- **LIME:** LimeTabularExplainer(training_data=np.array(X_train_smote), mode='classification', random_state=1)[2]

[1]https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html

[2]https://lime-ml.readthedocs.io/en/latest/lime.html#module-lime.lime_tabular

Resultados

Avaliação das métricas estudadas

Kaggle fraud detection					
	Árvore de decisão	**Regressão logística**	**Floresta aleatória**	**SVM**	**Redes neurais**
Acurácia balanceada	0.8851	**0.9443**	0.9245	0.9436	0.9032
Sensibilidade	0.7724	**0.9125**	0.8536	0.9064	0.8069
Especificidade	0.9978	0.9760	0.9953	0.9807	**0.9995**
Precisão	0.3782	0.0620	0.2487	0.0755	**0.7431**

IEEE-CIS fraud detection					
	Árvore de decisão	**Regressão logística**	**Floresta aleatória**	**SVM**	**Redes neurais**
Acurácia balanceada	0.5622	**0.6919**	0.6544	0.6644	0.6744
Sensibilidade	**0.6665**	0.6084	0.6021	0.4650	0.5441
Especificidade	0.4579	0.7753	0.7067	**0.8637**	0.8047
Precisão	0.0428	0.0894	0.0739	**0.1110**	0.1058

Credit card transaction					
	Árvore de decisão	**Regressão logística**	**Floresta aleatória**	**SVM**	**Redes neurais**
Acurácia balanceada	0.6829	0.8405	**0.8581**	0.7626	0.8125
Sensibilidade	0.3864	0.8008	0.8410	**0.8645**	0.8393
Especificidade	**0.9795**	0.8802	0.8751	0.6607	0.7858
Precisão	**0.0261**	0.0082	0.0083	0.0031	0.0059

Resultados

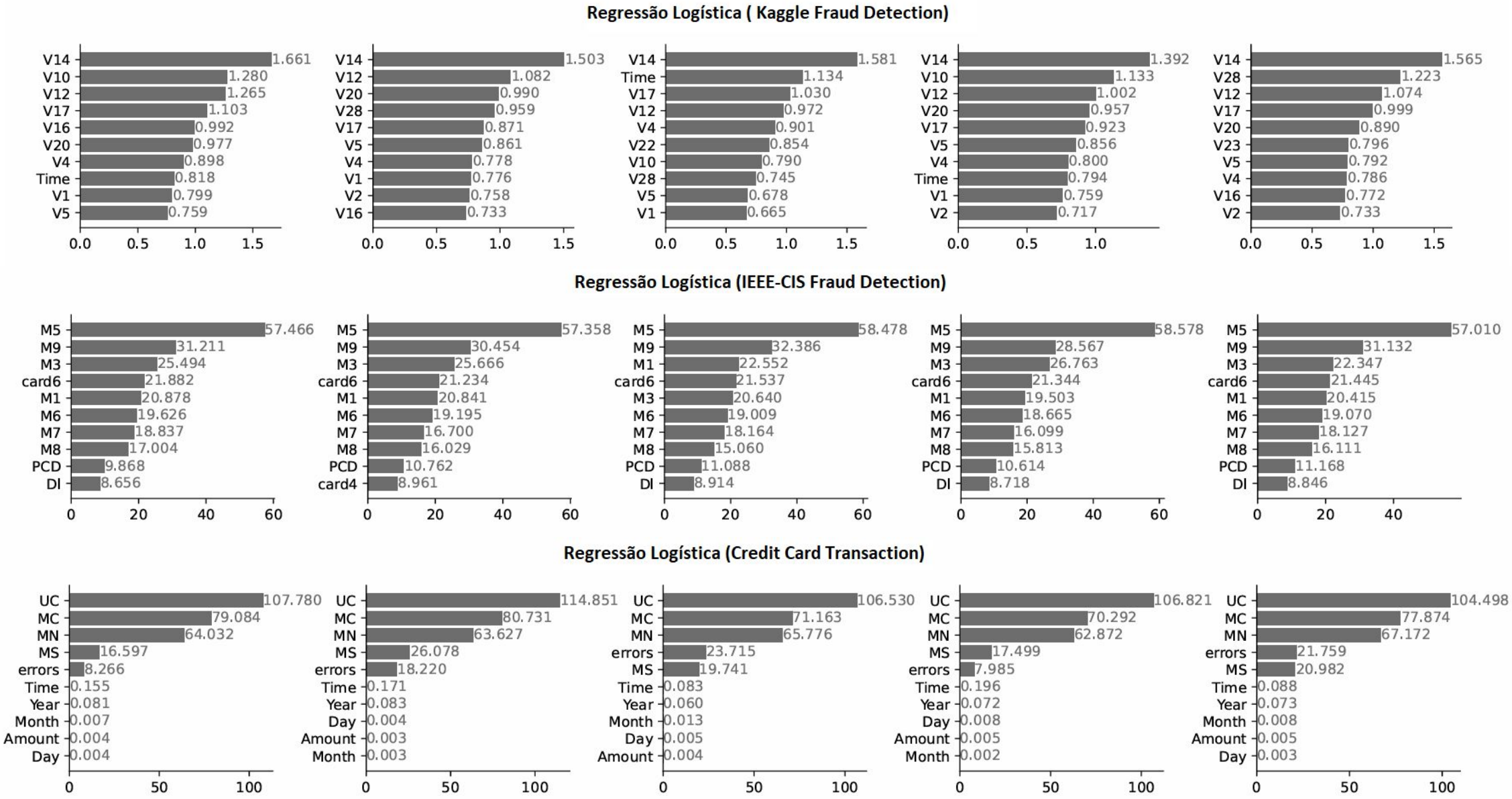
Avaliação das métricas estudadas

- Os modelos naturalmente interpretáveis não tiveram resultados muito inferiores aos modelos caixa-preta.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 11 2018.

Resultados

Pesos da regressão logística



Resultados

Tabela com agregado das concordâncias de SHAP e LIME com o proxy

	Kaggle Fraud Detection		IEEE-CIS Fraud Detection		Credit Card Transaction	
	SHAP	**LIME**	**SHAP**	**LIME**	**SHAP**	**LIME**
Árvore de decisão	21	**22**	**31**	23	**41**	40
Regressão logística	33	**35**	34	**35**	**41**	40
Florestas aleatórias	27	**30**	30	**33**	40	**41**
SVM	30	**35**	17	**21**	**40**	**40**
Rede neural	**31**	25	19	**20**	**40**	**40**

Resultados

Pesos da regressão logística

- Lime obteve resultados melhores que o SHAP na maior parte dos casos.
- O tempo de execução do método SHAP era maior.

Conclusões e trabalhos futuros

- SHAP e LIME podem ser boas alternativas para uma avaliação inicial das decisões de modelos;

- Os métodos de explicação avaliados podem ser sensíveis a variações nos dados utilizados;

- A aplicação de algoritmos naturalmente explicáveis em aplicações de detecção de fraudes com cartões de crédito pode ser uma alternativa viável.

Conclusões e trabalhos futuros

- Desbalanceamento dos conjuntos de dados utilizados no escopo do problema.
- Ajuste de hiperparâmetros.
- Como apresentar os resultados obtidos? **[Miller 2023]**
- Limitações de recursos computacionais

Observação

- Projeto de Graduação em Computação (PGC):
  - Comparativo de algoritmos de inteligência artificial explicável no mercado financeiro
  - Trabalho de conclusão apresentado no Bacharelado em Ciência da Computação da Universidade Federal do ABC (UFABC)
  - Orientado pelo professor Paulo Henrique Pisani

Referências

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., and Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7:93010–93022.

Chaudhary, K., Yadav, J., and Mallick, B. (2012). A review of fraud detection techniques: Credit card. International Journal of Computer Applications, 45:975–8887.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57:203–216.

Referências

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal L110, 59:1 88.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. http://arxiv.org/abs/1702.08608.

Alfaiz, N. S. and Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. Electronics (Switzerland), 11.

Referências

POZZOLO, A. D.; BORACCHI, G.; CAELEN, O.; ALIPPI, C. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, Institute of Electrical and Electronics Engineers Inc., v. 29, p. 3784–3797, 8 2018. ISSN 21622388;

Hsin, Y.-Y., Dai, T.-S., Ti, Y.-W., and Huang, M.-C. (2021). Interpretable electronic transfer fraud detection with expert feature constructions. In CIKM Workshops.

BOURDONNAYE, F. D. L.; DANIEL, F. Evaluating categorical encoding methods on a real credit card fraud detection database. 2021.

Referências

Moepya, S. O., Akhoury, S. S., Nelwamondo, F. V., and Twala, B. (2016). The role of imputation in detecting fraudulent financial reporting. International Journal of Innovative Computing, Information and Control ICIC International c, 12:333–356.

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., and Petitcolas, F. (2021). Explainable machine learning for fraud detection. Computer, 54(10):49–59.

Ji, Y. (2021). Explainable ai methods for credit card fraud detection: Evaluation of LIME and SHAP through a user study. https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva 20848.

Referências

Wu, T.-Y. and Wang, Y.-T. (2021). Locally interpretable one-class anomaly detection for credit card fraud detection. In 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), pages 25–30.

Chaquet-Ulldemolins, J., Gimeno-Blanes, F.-J., Moral-Rubio, S., Mu˜ noz-Romero, S., and Rojo-Álvarez, J.-L. (2022). On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable autoencoders. Applied Sciences, 12(8).

Referências

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”explaining the predictions of any classifier. volume 13-17-August-2016, pages 1135–1144. Association for Computing Machinery.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support. https://arxiv.org/pdf/2302.12389.

Contato

Gabriel Mendes
mendes.gabriel@aluno.ufabc.edu.br