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]

[1]Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC (UFABC) – Santo André – SP – Brasil

mendes.gabriel@aluno.ufabc.edu.br, paulo.pisani@ufabc.edu.br

Abstract. *Intelligent systems are used in the financial sector, including for fraud detection. In credit card transactions, machine learning algorithms can be used to obtain models which automate decisions such as classifying a transaction as fraudulent or not. In this context, this work presents a comparison between the explainable artificial intelligence techniques SHAP and LIME in models for fraud detection in credit card transactions, showing that these techniques can be suitable for the problem. The use of interpretable algorithms in critical sectors such as financial sector is also discussed, as well as the effectiveness and need for explainable artificial intelligence techniques.*

Resumo. *Sistemas inteligentes são utilizados no mercado financeiro, inclusive para detecção de fraudes. Em transações com cartões de crédito, algoritmos de aprendizado de máquina podem ser usados para obter modelos que automatizam decisões como classificar uma transação como fraudulenta ou não. Neste contexto, este trabalho apresenta uma comparação entre as técnicas de inteligência artificial explicável SHAP e LIME em modelos para detecção de fraudes em transações com cartão crédito, mostrando que essas técnicas podem ser adequadas ao problema. Também é discutida a utilização de algoritmos naturalmente explicáveis, assim como a efetividade e a necessidade de técnicas de inteligência artificial explicável.*

1. Introdução

Modelos de aprendizado de máquina foram utilizados para detecção de fraudes em transações com cartões de crédito [Makki et al. 2019, Chaudhary et al. 2012]. Contudo, diversos modelos obtidos com aprendizado de máquina não são facilmente explicáveis, ou seja, o seu processo de decisão não é facilmente compreensível por seres humanos. Ao utilizar aprendizado automático em ambientes complexos, como no caso do mercado financeiro, é importante que os modelos obtidos tenham bom desempenho preditivo e também tenham um funcionamento seguro. A segurança pode ser alcançada por meio da interpretabilidade do modelo de aprendizado automático [Doshi-Velez and Kim 2017]. Nesse contexto, interpretabilidade pode ser definida como o grau de entendimento que um ser humano tem das decisões tomadas por um modelo [Molnar 2022].

No mercado financeiro, diversas decisões podem ser tomadas por modelos obtidos com aprendizado automatizado, como aprovação ou não de um empréstimo, aceitação ou não de uma transação de cartão de crédito, etc. Entretanto, o uso de modelos que tomam

decisões que não podem ser explicadas não são adequados para sistemas altamente regulados, como o setor bancário [Bussmann et al. 2021]. É necessário que as decisões tomadas pelos modelos tenham explicações que possam ser compreendidas por seres humanos. A União Europeia e seus países membros exigem isso por lei [European Union 2016]. Somado a isso, também há questões éticas envolvidas. Sistemas inteligentes devem ser responsáveis por suas decisões, além de serem auditáveis [Bussmann et al. 2021].

Nesse contexto, inteligência artificial explicável pode trazer contribuições importantes. Técnicas e modelos que tornam o comportamento e as predições de modelos de aprendizado de máquina compreensíveis para seres humanos fazem parte do que é conhecido como inteligência artificial explicável [Molnar 2022]. A proposta de novas técnicas e avanços nas existentes, juntamente com o crescimento da área de inteligência artificial explicável [Vilone and Longo 2020], motiva a exploração da aplicação dessas técnicas em diversos setores que envolvem inteligência artificial.

Durante a pandemia de COVID-19, serviços financeiros online passaram a ser uma necessidade [Psychoula et al. 2021]. Em 2019, a quantidade de dinheiro movimentado por fraudes financeiras em ambientes online era 80% maior que o PIB do Reino Unido [Gee et al. 2019]. Nesse contexto, formas de mitigar esse problema são necessárias. O uso de aprendizagem automática com técnicas de inteligência artificial explicável pode contribuir para a auditoria dos modelos obtidos, assim como para uma melhor compreensão dos resultados obtidos [Hanif 2021].

Uma aplicação importante no contexto do mercado financeiro é a detecção de fraudes em transações com cartões de crédito. Fraude neste contexto refere-se à utilização de cartões de crédito para obtenção de produtos ou serviços de forma ilegal [Chaudhary et al. 2012]. Neste trabalho, o problema tratado é o de, com base em um histórico de transações realizadas com cartões de crédito já classificadas como fraude ou legítimas, obter modelos que possam classificar outras transações como fraudulentas ou legítimas. Algumas técnicas de inteligência artificial explicável aplicadas em trabalhos anteriores no contexto de transações com cartões de créditos são o LIME (*Local Interpretable Model-agnostic Explanations*) e o SHAP (*Shapley Additive Explanations*) [Ji 2021, Psychoula et al. 2021].

Este trabalho tem o objetivo de comparar técnicas de inteligência artificial explicável em modelos para detecção de fraudes em cartões de crédito utilizando aprendizado de máquina. Como objetivos específicos, o trabalho envolveu selecionar conjuntos de dados e algoritmos de classificação usados para detecção de fraudes com cartões de crédito, escolher técnicas de inteligência artificial explicável e realizar o comparativo entre essas técnicas. A pesquisa e os resultados apresentados neste artigo são parte do trabalho de conclusão de curso Projeto de Graduação em Computação (PGC) *Comparativo de algoritmos de inteligência artificial explicável no mercado financeiro* apresentado no curso Bacharelado em Ciência da Computação da Universidade Federal do ABC (UFABC). O restante do texto está organizado da seguinte forma: na Seção 2, serão introduzidos alguns trabalhos relacionados; na Seção 3, será explicada brevemente a metodologia experimental adotada neste trabalho, incluindo conjuntos de dados e configuração dos experimentos; na Seção 4, os resultados obtidos são discutidos, incluindo uma discussão sobre a efetividade e necessidade da utilização de técnicas de inteligência artificial explicável; e, na Seção 5, são apresentadas as conclusões finais e trabalhos futuros.

2. Trabalhos relacionados

Diversos trabalhos exploraram a aplicação de técnicas de inteligência artificial explicável no problema de detecção de fraudes em transações com cartão de crédito.

Em [Ji 2021], foi realizado um estudo sobre a eficiência da aplicação de LIME e de SHAP para obter explicações dos resultados de seus modelos. De acordo com os autores, a confiança dos usuários no sistema aumentava com o fornecimento das explicações de suas decisões, sendo que LIME obteve um desempenho um pouco superior ao SHAP na avaliação realizada. Os algoritmos LIME e SHAP também foram avaliados no trabalho de [Psychoula et al. 2021]. Nesse trabalho, foi investigada a aplicação dessas técnicas de inteligência artificial explicável para a detecção de fraudes com cartões de crédito em tempo real com modelos de aprendizado supervisionado e não supervisionado. No geral, a conclusão apresentada pelos autores foi de que LIME e SHAP podem ser técnicas adequadas para obter explicações de modelos usados para detecção de fraudes financeiras.

Outro trabalho que estudou detecção de fraudes com cartões de crédito foi [Hsin et al. 2021]. Os autores argumentaram que abordagens tradicionais baseadas em regras para detecção de fraudes com cartões de crédito não são efetivas para este problema. No trabalho, foi proposta a utilização de preditores baseados em comportamento e segmentação obtidos a partir de contas não fraudulentas. Ao utilizar esses preditores, podem ser obtidos melhores resultados em modelos baseados em árvores, que podem ser considerados modelos naturalmente interpretáveis.

Um *framework* para detecção de fraudes com cartões de crédito foi proposto por [Wu and Wang 2021]. A proposta envolveu um módulo de interpretabilidade, que é responsável por gerar explicações para as predições. A detecção de fraudes foi realizada por meio de duas redes neurais profundas treinadas de forma adversarial e sem supervisão. O módulo de interpretabilidade possuía três explicadores naturalmente interpretáveis.

O trabalho de [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*. A metodologia proposta possui duas vantagens principais. Primeiro, essa metodologia pode ser aplicada em conjunto com qualquer modelo de aprendizado de máquina. Além disso, a nova metodologia permite mapear as entradas com as saídas dos modelos, contribuindo para o cumprimento de requisitos regulatórios de interpretabilidade na detecção de fraudes com cartões de crédito.

Diferentemente da maioria dos trabalhos citados nesta seção que consideram apenas um conjunto de dados, este trabalho faz uma avaliação de técnicas de inteligência artificial explicável com três conjuntos de dados. Cada conjunto possui diferentes condições de desbalanceamento. O trabalho de [Chaquet-Ulldemolins et al. 2022] considerou três conjuntos de dados também, mas diferentes dos usados neste artigo. Além da avaliação prática dos resultados das técnicas de inteligência artificial explicável, este artigo inclui uma discussão sobre a efetividade e necessidade da utilização dessas técnicas.

3. Metodologia experimental

Esta seção apresenta a metodologia experimental adotada, incluindo conjuntos de dados, separação dos dados e algoritmos de classificação. Neste trabalho[1], foi realizado um

[1] `https://github.com/GaMendes/pgc`

comparativo de técnicas de inteligência artificial explicável com modelos de aprendizado de máquina aplicados na detecção de fraudes em transações financeiras com cartões de crédito. Nos experimentos, foram considerados algoritmos que não são naturalmente explicáveis, assim como algoritmos que são naturalmente explicáveis.

3.1. Conjunto de dados

Foram utilizados conjuntos de dados com transações feitas com cartões de crédito nos experimentos. Cada transação pode ser classificada como fraudulenta ou legítima (não fraudulenta). Neste trabalho, a classe positiva foi definida como a classe que representa transações fraudulentas. Neste contexto, três conjuntos de dados foram selecionados:

- *Kaggle fraud detection*[2]: Esse conjunto de dados contém transações feitas por cartões de crédito em setembro de 2013 por titulares de cartões europeus. Nesse conjunto de dados, há 284.807 transações que ocorreram durante dois dias. Dessas transações, apenas 492 são fraudes (0,172%), ou seja, há um grande desbalanceamento entre as classes. No total, há 30 preditores, sendo que 28 foram obtidos por meio de PCA (*Principal Component Analisys*) e os outros dois (tempo e valor) não passaram por essa transformação. Esses dois últimos atributos passaram por uma reescala usando RobustScaler do *scikit-learn* [Pedregosa et al. 2011]. Diversos trabalhos utilizaram esse conjunto de dados, como [Dal Pozzolo et al. 2014], [Pozzolo et al. 2015], [Dal Pozzolo et al. 2017], [Carcillo et al. 2017], [Carcillo et al. 2019] e [Le Borgne et al. 2022].
- *IEEE-CIS fraud detection*[3]: Esse conjunto de dados foi utilizado em uma competição de detecção de fraudes realizada no ano de 2019. O *IEEE-CIS fraud detection* contém dados de transações com cartões de crédito e informações dos clientes obtidas por uma empresa de pagamentos. Assim no caso do conjunto de dados do *Kaggle*, esse também é altamente desbalanceado, sendo que apenas 3,49% das transações são fraudes. No total, o conjunto de dados tem 433 preditores. Assim como realizado no trabalho de [Psychoula et al. 2021], não foram utilizados todos os preditores. A maioria dos preditores não contém uma descrição detalhada de seus valores. Além disso, em razão de limitações de recursos computacionais disponíveis para realizar os experimentos, foram utilizados apenas 19 desses preditores nos experimentos. Alguns registros possuíam dados com valores não preenchidos e, nesses casos, foi utilizado *imputation* [Moepya et al. 2016]. Preditores categóricos foram transformados usando CatBoost [Bourdonnaye and Daniel 2021]. Originalmente, os dados são divididos em subconjuntos para treino e para teste. Entretanto, este trabalho adota uma metodologia diferente para separação dos dados, conforme descrito na próxima seção. Apenas o subconjunto de treino (cerca de 500 mil registros) foi usado, sendo posteriormente dividido em treino e teste nos experimentos. O percentual de transações fraudulentas nesse subconjunto é de 3,50%.
- *Credit card transaction*[4]: Esse conjunto de dados contém 2,4 milhões de registros, com um total de 12 preditores, gerados sinteticamente [Padhi et al. 2021]. As transações foram geradas por amostragem estocástica [Padhi et al. 2021]. Das

[2]`https://www.kaggle.com/mlg-ulb/creditcardfraud`
[3]`https://kaggle.com/competitions/ieee-fraud-detection`
[4]`https://github.com/IBM/TabFormer`

2,4 milhões de transações registradas, apenas 29.342 são fraudulentas (0,012%). Nos experimentos, foram utilizados somente 480.000 registros desse conjunto, mas mantendo o mesmo percentual de transações fraudulentas. Assim como no caso do conjunto de dados anterior, dados não preenchidos passaram por *imputation* [Moepya et al. 2016] e preditores categóricos foram transformados usando CatBoost [Bourdonnaye and Daniel 2021].

3.2. Separação dos dados

Os dados foram separados entre treino e teste utilizando validação cruzada com o *k-fold* (k = 5) estratificado [Alfaiz and Fati 2022]. Dessa forma, os dados foram separados em cinco subconjuntos, sendo quatro usados para treinamento e um para teste. O processo foi repetido por cinco iterações, sendo que, a cada iteração, um subconjunto diferente foi usado para teste.

Além disso, como há um grande desbalanceamento entre as classes positiva (transação fraudulenta) e negativa (transação não fraudulenta), foi utilizado SMOTE sobre os dados balanceando a quantidade de exemplos de cada classe no conjunto de treinamento. Na literatura, o método SMOTE foi utilizado em trabalhos anteriores que lidaram com fraudes com cartões de crédito [Alfaiz and Fati 2022].

3.3. Algoritmos de classificação

Nos experimentos, foram avaliados os algoritmos de classificação regressão logística, árvore de decisão, florestas aleatórias, SVM e rede neural artificial usando as implementações disponíveis no *scikit-learn* [Pedregosa et al. 2011]. Neste trabalho, regressão logística, árvore de decisão e florestas aleatórias foram considerados como algoritmos de classificação naturalmente explicáveis. Por outro lado, SVM e rede neural artificial foram considerados como algoritmos de classificação que não são naturalmente explicáveis. A escolha dos algoritmos de classificação foi realizada com base na literatura pertinente no contexto do problema.

Algumas configurações adotadas para os algoritmos são apresentadas a seguir:

- Regressão logística: foi utilizado *max_iter* = 2000 e *solver* = "newton-cholesky", assim como recomendado na documentação do *scikit-learn* para conjuntos de dados com muito mais registros do que colunas[5];
- Florestas aleatórias: a propriedade *max_depth* assumiu o valor 2;
- Árvore de decisão e redes neurais artificiais: foi usada a configuração padrão do *scikit-learn*. A rede neural foi criada utilizando a classe *MLPClassifier*[6].
- SVM: foi usado o *kernel* linear (LinerSVC), *max_iter* igual a 2000 e a propriedade *dual* como falso, como sugere a documentação do *scikit-learn* [7] para os tipos de conjunto de dados utilizados nos experimentos;

Ao longo dos experimentos, a semente pseudo-aleatória (*random_state*) dos algoritmos regressão logística, árvore de decisão, florestas aleatórias, redes neurais e SVM assumiu o valor 1.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

3.4. Técnicas de inteligência artificial explicável

Nos experimentos, duas técnicas de inteligência artificial explicável foram comparadas: SHAP e LIME.

SHAP é uma abordagem unificada para tentar explicar a saída de modelos de aprendizado automático utilizando o conceito de valores *Shapley* da teoria dos jogos [Lundberg and Lee 2017]. A técnica aproxima o cálculo de valores *Shapley* a partir do que os criadores da técnica chamaram de preditores aditivos. Para os experimentos apresentados na próxima seção, SHAP foi implementado utilizando uma biblioteca disponibilizada por um dos criadores da técnica SHAP. Os explicadores baseados em SHAP foram criados utilizando a classe *Explainer*[8] da biblioteca SHAP, com o conjunto de treino transformado com SMOTE como conjunto de *background*.

A outra técnica comparada foi o LIME. Essa técnica pode explicar as predições de qualquer classificador aproximando esta predição localmente utilizando um modelo de aprendizado automático interpretável [Ribeiro et al. 2016]. Nos experimentos realizados, foi utilizado o método LIME disponibilizado pela classe *LimeTabularExplainer*[9], com a propriedade *mode* em classificação e a propriedade *training_data* igual ao conjunto de treino.

4. Resultados e discussão

Os experimentos foram realizados utilizando algoritmos comumente aplicados ao problema de detecção de fraudes em transações com cartões de crédito para que, posteriormente, fosse possível avaliar a contribuição individual dos preditores dos diferentes conjuntos de dados, em cada um dos modelos criados. Essa avaliação foi realizada utilizando SHAP e LIME. A partir dos resultados das técnicas de inteligência artificial explicável, foi realizado um comparativo das explicações obtidas.

4.1. Desempenho geral

A Tabela 1 apresenta as médias de algumas métricas de avaliação após a validação cruzada descrita na Seção 3.2 para os três conjuntos de dados considerados.

Analisando os dados da Tabela 1, é possível verificar que, de maneira geral, os valores de precisão obtidos foram baixos. Isso sugere que os classificadores não foram capazes de fazer previsões precisas para a classe positiva (fraude). Um fator que pode ter contribuído para esse resultado é o alto desbalanceamento entre as classes nos conjuntos de dados. Além disso, um melhor ajuste de hiperparâmetros dos algoritmos também poderia contribuir para melhores resultados preditivos.

Os modelos baseados em redes neurais e SVM dominaram as métricas de precisão e especificidade nos dois primeiros conjuntos de dados. No entanto, em alguns casos, é possível observar que os modelos naturalmente interpretáveis, como regressão logística e árvore de decisão, obtiveram resultados que não foram muito inferiores a algoritmos que não são naturalmente interpretáveis, como redes neurais e SVM. No terceiro conjunto de dados, o resultado para precisão e especificidade foi inclusive maior para a árvore de

[8] `https://shap-lrjball.readthedocs.io/en/latest/generated/shap.Explainer.html`

[9] `https://lime-ml.readthedocs.io/en/latest/lime.html`

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

Tabela 1. Desempenho preditivo dos modelos criados a partir dos cinco algoritmos de classificação e dos três conjuntos de dados. O melhor resultado para cada métrica em cada conjunto de dados foi destacado em negrito.

decisão. Isso sugere que modelos naturalmente interpretáveis podem ser uma opção adequada para sistemas de detecção de fraudes em transações com cartões de crédito. Na literatura, o trabalho de [Rudin 2019] argumenta que sistemas críticos devem ser naturalmente explicáveis, evitando o uso de modelos caixa-preta nesse tipo de aplicação.

4.2. Explicações locais

Uma comparação entre as técnicas de inteligência artificial explicável SHAP e LIME é apresentada nesta seção. Os pesos atribuídos pela regressão logística, que é um modelo naturalmente explicável, foram utilizados para obter uma lista com os preditores mais importantes em cada conjunto de dados, similar ao *ranking* discutido em [Psychoula et al. 2021]. Essa lista foi usada como referência para avaliar os resultados das técnicas SHAP e LIME ao realizar a explicação dos exemplos selecionados. Essa metodologia corresponde a realizar uma explicação por função, definindo os pesos da regressão logística como *proxy* [Doshi-Velez and Kim 2017].

A Figura 1 contém os 10 preditores mais importantes para cada conjunto de dados de acordo com os pesos atribuídos pela regressão logística. É importante destacar que o primeiro conjunto de dados contém preditores resultado de uma transformação com PCA. Isso torna difícil obter explicações para pessoas sem conhecimento do processo de codificação. Contudo, é importante avaliar como técnicas de inteligência artificial explicável se comportam em conjuntos de dados transformados, uma situação comum em relação a conjuntos de dados dentro do escopo do problema de detecção de fraudes com cartões de crédito [Psychoula et al. 2021].

Após a obtenção dos 10 preditores mais importantes pela regressão logística, foi realizada a avaliação das técnicas de inteligência artificial explicável. Para isso, foram realizadas explicações locais com essas duas técnicas a partir de exemplos da classe positiva (fraude) selecionados aleatoriamente a partir dos dados de treino. Foi selecionado um exemplo em cada uma das iterações da validação cruzada com *k-fold*. Explicadores foram

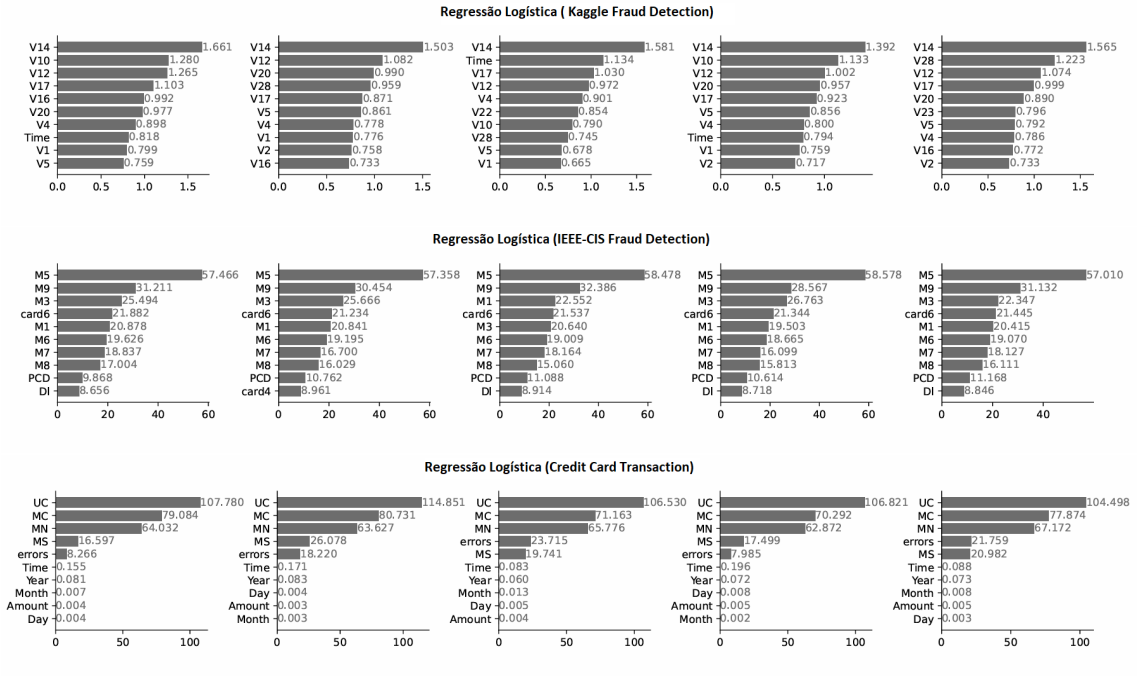

Figura 1. Lista dos 10 preditores mais importantes considerando o valor absoluto dos pesos atribuídos pela regressão logística na obtenção dos modelos (os pesos foram calculados de maneira global). Cada gráfico representa uma iteração diferente da validação cruzada ($k = 5$) para cada um dos três conjuntos de dados.

criados usando dados de treino, pois o objetivo foi analisar os preditores mais importantes aos modelos avaliados [Molnar 2022]. A partir dos modelos obtidos, foram elencados os preditores mais importantes. Para a avaliação realizada, foram considerados os valores *Shapley* absolutos aproximados pela técnica SHAP, e os valores absolutos retornados pela técnica LIME. As duas técnicas foram aplicadas de maneira local usando um exemplo aleatório da classe positiva (fraude) usado no treinamento, conforme descrito no início desta seção.

A partir das listas dos preditores mais importantes obtidos pela regressão logística e pelas técnicas SHAP e LIME, foi realizado um comparativo. Para isso, foi calculada a quantidade de preditores na lista dos 10 mais importantes de cada técnica que apareceu na lista dos 10 mais importantes da regressão logística. A ordem em que cada preditor aparece não foi considerada para essa avaliação. Esse cálculo foi realizado para cada exemplo selecionado, ou seja, foi realizado uma vez para cada iteração da validação cruzada *k-fold*.

A Tabela 2 apresenta o resultado da comparação entre as listas de 10 preditores obtidas por SHAP e por LIME. O resultado é um somatório das quantidades obtidas em cada iteração da validação cruzada, conforme descrito no parágrafo anterior. Como foram cinco exemplos considerados e, para cada exemplo, a quantidade máxima de preditores comuns é 10, o valor máximo para o somatório é 50 nesta tabela.

	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**

Tabela 2. Comparativo dos métodos SHAP e LIME para os três conjuntos de dados. A avaliação foi realizada de acordo com número de vezes que os 10 preditores mais importantes dos métodos LIME e SHAP aparecem na lista dos preditores mais importantes calculados pelos pesos da regressão logística em suas respectivas iterações da validação cruzada (a ordem em que cada atributo ocorre nas listas não foi considerada). Os melhores resultados foram destacados em negrito.

Observando os valores obtidos, tanto LIME quanto SHAP obtiveram resultados similares, sendo que LIME obteve resultados um pouco melhores na avaliação realizada. Uma diferença importante observada na execução dos experimentos foi que as explicações geradas com SHAP consumiram um tempo maior para serem obtidas. Todos os experimentos foram realizados utilizando o mesmo computador. Um outro ponto que pode ser considerado é que, para as duas técnicas, ocorreram grandes variações na ordem dos preditores em relação a cada iteração da validação cruzada. Isso pode indicar que essas técnicas são suscetíveis a perturbações [Alvarez-Melis and Jaakkola 2018], visto que, a cada iteração da validação cruzada, os dados de treino e de teste eram diferentes.

4.3. Discutindo as explicações

Um aspecto importante a ser considerado ao utilizar inteligência artificial explicável é avaliar a utilidade das explicações obtidas, no sentido de entender o escopo do problema avaliado em relação ao desempenho do modelo. As explicações com as duas técnicas, como foram aplicadas neste trabalho, podem não ser adequadas para pessoas tomando decisões sobre o problema. Apenas mostrar os valores *Shapley* retornados pela técnica SHAP ou os valores retornados pela técnica LIME pode não ser suficiente. É provável que apenas os desenvolvedores dos sistemas de explicação consigam tirar informações úteis para a tomada de decisões com base nos resultados obtidos [Miller et al. 2017].

Para que a aplicação de métodos de inteligência artificial explicável seja efetiva no momento de solucionar problemas, as explicações geradas devem ser úteis a pessoas que não tem um alto grau de conhecimento da construção do sistema ou dos métodos utilizados, isto é, as explicações tem de ser feitas pensando no usuário que vai recebê-las [Miller et al. 2017]. Explicações geradas pensando no usuário são mais fáceis de entender, e geram informações úteis para a avaliação dos resultados, mas há ainda o problema de que usuários não interagem com explicações o suficiente para ajudar no processo de decisão baseada nas recomendações dos modelos, o que pode acontecer por diversas razões [Miller 2023].

Um novo paradigma para lidar com esse problema, a inteligência artificial avaliativa, pode resolver esse problema [Miller 2023]. Nesse paradigma, as explicações, ao invés de fornecerem apenas informações para que o usuário aceite ou não a saída do modelo, fornece evidências para diversos cursos de ações diferentes. Essa forma, além de ser

mais alinhada com o processo cognitivo de tomada de decisão, pode contribuir para que os usuários do sistema questionem as decisões tomadas [Miller 2023].

5. Conclusão

Este trabalho apresentou uma comparação sobre a aplicação das técnicas de inteligência artificial explicável SHAP e LIME no problema de detecção de fraudes em transações com cartões de crédito. Essas técnicas foram aplicadas sobre modelos obtidos a partir de cinco algoritmos de classificação, que foram avaliados em três conjuntos de dados. Apesar de possuírem diferentes condições de desbalanceamento, todos os conjuntos possuíam um alto grau de desbalanceamento.

No geral, os resultados das métricas de predição indicam que algoritmos naturalmente explicáveis também podem ser uma alternativa a algoritmos não naturalmente explicáveis dentro do contexto do problema. Isso pode contribuir para atender à argumentação de [Rudin 2019] de que sistemas críticos devem ser naturalmente explicáveis. Sobre às técnicas LIME e SHAP, foi realizada uma avaliação comparando o desempenho de acerto dos 10 preditores mais importantes com relação à lista de preditores mais importantes obtida a partir dos pesos atribuídos pela regressão logística.

O trabalho considerou apenas explicações locais nas avaliações realizadas. Em trabalhos futuros, explicações globais podem ser realizadas, como a avaliação do método SHAP agregando os valores de suas explicações locais em uma única explicação para todas as instâncias. Isso permitiria observar como essa técnica se compararia à avaliação dos pesos globais da regressão logística. Uma outra vantagem que explicações globais trariam seria que os resultados teriam uma menor probabilidade de ficarem enviesados a depender da escolha aleatória de um exemplo local. Um estudo sobre a representatividade dos dados de treino e teste também pode ser feito no futuro. Além disso, trabalhos futuros podem considerar um estudo de técnicas de ajuste de hiper-parâmetros, o que pode melhorar o desempenho preditivo dos algoritmos de classificação e, possivelmente, impactar nos resultados obtidos pelas técnicas de inteligência artificial explicável.

Outra alternativa para pesquisas futuras no contexto do problema estudado neste trabalho seria avaliar maneiras de apresentar os resultados obtidos. Considerando que os conjuntos de dados utilizados no problema contém informações confidenciais e frequentemente passam por processo de transformação, tornar-se um desafio obter explicações úteis e confiáveis. Isso se estende à criação dos modelos pois, quando um preditor passou por um processo de transformação, pode ser difícil avaliar sua contribuição na resolução do problema. Uma avaliação dos conjuntos de dados de maneira unificada, considerando preditores comuns também pode ser realizada em trabalhos futuros.

Referências

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

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. http://arxiv.org/abs/1806.08049.

Bourdonnaye, F. D. L. and Daniel, F. (2021). Evaluating categorical encoding methods on a real credit card fraud detection database. https://arxiv.org/pdf/2112.12024.

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

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., and Bontempi, G. (2017). Scarff : a scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., and Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*.

Chaquet-Ulldemolins, J., Gimeno-Blanes, F.-J., Moral-Rubio, S., Muñoz-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).

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

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., and Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–14.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41:4915–4928.

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

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.

Gee, J., Button, M., and Brooks, G. (2019). *The financial cost of fraud: what data from around the world shows*. MacIntyre Hudson. Institution: University of Portsmouth. Department: Institute of Criminal Justice Studies.

Hanif, A. (2021). Towards explainable artificial intelligence in banking and financial services. http://arxiv.org/abs/2112.08441.

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*.

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.

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.

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.

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.

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

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum. https://arxiv.org/pdf/1712.00547.

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.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition. https://christophm.github.io/interpretable-ml-book.

Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., and Altman, E. (2021). Tabular transformers for modeling multivariate time series. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3565–3569. IEEE.

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.

Pozzolo, A. D., Caelen, O., Johnson, R. A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166.

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.

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.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. http://arxiv.org/abs/2006.00093.

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.