

# Avaliação de Técnicas de Balanceamento de Dados na Detecção de Fraude em Transações Online de Cartão de Crédito\*

Arthur Cavalcanti<sup>1</sup>, Diego Brandão<sup>1</sup>, Eduardo Bezerra<sup>1</sup>, Rafaelli Coutinho<sup>1</sup>

<sup>1</sup>Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)

arthur.silveira@aluno.cefet-rj.br, diego.brandao@cefet-rj.br,

rafaelli.coutinho@cefet-rj.br, ebezerra@cefet-rj.br

**Abstract.** *Due to the increase in e-commerce and credit card use, fraud detection has become a major challenge for the entities involved. Despite the losses, this type of fraud still represents a small portion of transactions, creating a data imbalance problem in the areas of fraud detection within the financial system. This work evaluates various combinations of feature selection techniques, class balancing, and classification. Undersampling, oversampling techniques, and classifier threshold adjustments were used to balance the classes. The combinations were tested on two imbalanced datasets and evaluated using the  $F_1$ -score metric. The results show improved performance when implementing data balancing techniques and optimal classification threshold adjustments.*

**Resumo.** *Devido ao aumento do comércio eletrônico e do uso de cartões de crédito, as fraudes com cartões de crédito tornaram-se um grande desafio para as entidades envolvidas. Apesar dos prejuízos, essas fraudes ainda representam uma pequena parte das transações, criando um problema de desbalanceamento de dados nas áreas de detecção de fraudes do sistema financeiro. Este trabalho avalia várias combinações de técnicas de seleção de atributos, balanceamento de classes e algoritmos de classificação. Para balancear as classes, foram usadas técnicas de subamostragem, superamostragem e ajustes de limiares nos classificadores. As combinações foram testadas em dois conjuntos de dados desbalanceados, avaliados pela métrica  $F_1$ . Os resultados mostram um ganho de desempenho quando são implementadas técnicas de balanceamento de dados e otimização de limiares de classificação.*

## 1. Introdução

Com o amplo acesso à internet e expansão da pandemia de coronavírus em 2020, o uso de cartões de crédito para compras online cresceu significativamente. No entanto, esse aumento também elevou a incidência de fraudes, causando prejuízos para consumidores, varejistas e instituições financeiras: segundo o *Global Fraud Report*<sup>1</sup> de 2022 da CyberSource<sup>2</sup>, a receita perdida por fraudes aumentou de 3,1% em 2021 para 3,6% em 2022.

\*Os autores agradecem ao CNPq, CAPES, FAPERJ e IIA pelo patrocínio parcial desta pesquisa.

<sup>1</sup><https://www.cybersource.com/content/dam/documents/campaign/fraud-report/global-fraud-report-2022.pdf>

<sup>2</sup>Empresa especializada em prevenção à fraude e uma das referências no assunto.

Esse cenário impulsiona a busca por métodos mais eficazes de prevenção e detecção de fraudes.

Karthika and Senthilselvi [2023] e Ghaleb et al. [2023] indicam que a detecção de fraudes em transações online é um desafio sobretudo devido ao desbalanceamento de classes, onde o número de transações fraudulentas é bem menor que o número de transações genuínas [Ileberi et al., 2021]. Decorrente desse desbalanceamento, Leevy et al. [2023] citam a importância da validação cruzada com estratificação durante a classificação para melhorar a capacidade de generalização, assim como Muaz et al. [2020] citam o mesmo benefício para a estratificação durante a separação dos conjuntos de treino e teste.

Ileberi et al. [2021], Priscilla and Prabha [2020] e Zhang et al. [2019] demonstram a importância de aplicar as soluções de balanceamento e detecção de fraudes em mais de um conjunto de dados para maior generalização. Ileberi et al. [2021], Priscilla and Prabha [2020], Zhang et al. [2019], Muaz et al. [2020] e Xie et al. [2021] mostram a importância de utilizar diversos classificadores em modelos de detecção de fraudes para avaliação de desempenho. Leevy et al. [2023] e Prabha and Priscilla [2024] utilizam a busca por limiares ótimos de classificação como uma técnica de balanceamento de dados. Alguns trabalhos também realizam análises comparativas e até inovadoras de técnicas de balanceamento com o intuito de avaliar quais técnicas são mais eficazes [Makki et al., 2019; Sisodia et al., 2017; Gupta et al., 2023; Priscilla and Prabha, 2020; Muaz et al., 2020; Bhagwani et al., 2021]; no entanto, não exploram todos os pontos importantes citados nos outros trabalhos.

Este trabalho busca cobrir esses pontos, comparando diferentes técnicas de balanceamento de dados e avaliando seu desempenho em diferentes cenários. Tais cenários combinam métodos de seleção de atributos, técnicas de balanceamento de dados e algoritmos de classificação. Para tanto, dois conjuntos de dados de transações de cartões de crédito altamente desbalanceados foram utilizados, aplicando estratificação de dados de treino e teste e validação cruzada para garantir a eficácia dos modelos de aprendizado de máquina. A contribuição desse trabalho está na diversidade de técnicas avaliadas em diferentes bases de dados, focando em métricas de desempenho específicas para transações fraudulentas.

## 2. Referencial Teórico

No contexto de aprendizado de máquina, técnicas como a seleção de atributos [Laborda and Ryoo, 2021] e técnicas para tratamento dos dados (*e.g. one-hot encoding*) são usadas para reduzir a complexidade dos dados e melhorar o desempenho dos modelos [Jahnavi et al., 2023]. Também nesse contexto, técnicas de balanceamento de classes, como superamostragem e subamostragem, são essenciais para melhorar a representatividade das classes e o desempenho dos modelos [Sun et al., 2009; He and Garcia, 2009].

A superamostragem aumenta a proporção da classe minoritária criando novas instâncias sintéticas [He and Garcia, 2009] e inclui técnicas como *Resample*, *Random OverSampling* (ROS), *Synthetic Minority Oversampling Technique* (SMOTE), *Borderline Synthetic Minority Oversampling Technique* (BSMOTE) e *Adaptive Synthetic Sampling* (ADASYN). Por outro lado, a subamostragem reduz a quantidade de instâncias da classe majoritária para alcançar o equilíbrio e inclui técnicas como *Resample*, *Random Under-sampling* (RUS), *Near Miss* e *Instance Hardness Threshold* (IHT) [Hasib et al., 2020].

Técnicas de detecção de fraudes de cartão de crédito são classificadas em aprendizado supervisionado e não supervisionado. No aprendizado não supervisionado, técnicas de agrupamento são usadas para identificar comportamentos anômalos sem a necessidade de rotulação prévia [Carcillo et al., 2021]. No aprendizado supervisionado, técnicas como *Extremely Randomized Trees* (ET), *Random Forest* (RF), *Árvores de Decisão* (AD), *CatBoost* (CB), *Gradient Boosting* (GB), *AdaBoost* (AB), *Regressão Logística* (RL), *K-Nearest Neighbors* (KNN) e *Naive Bayes* (NB) são aplicadas a conjuntos de dados rotulados para prever transações fraudulentas [Bhattacharyya et al., 2011].

Vale ressaltar que, no momento da classificação, a validação cruzada revela-se uma técnica estatística útil, avaliando modelos diversas vezes para detectar a presença de sobreajuste - ou subajuste - e a capacidade preditiva deles [Amit Singh and Tiwari, 2022]. Outro ponto importante durante a classificação é o limiar de decisão - o valor que determina a qual classe uma entrada será atribuída (por exemplo, fraude ou não fraude). Nesse sentido, a busca por um limiar que equilibra as taxas de falsos positivos e falsos negativos e otimiza o modelo pode funcionar como fator de ajuste em conjuntos de dados desbalanceados e, conseqüentemente, como uma técnica de balanceamento de classes [Leevy et al., 2023]. O processo envolve, então, experimentar diferentes valores de limiares e avaliar o desempenho do modelo em cada configuração.

### 3. Metodologia

As etapas que compõem a metodologia adotada estão ilustradas na Figura 1 e consistem em: 1) tratamento de dados, 2) divisão dos dados (conjunto de treino e teste), 3) balanceamento de classes, 4) seleção de atributos, 5) classificação, e 6) avaliação.

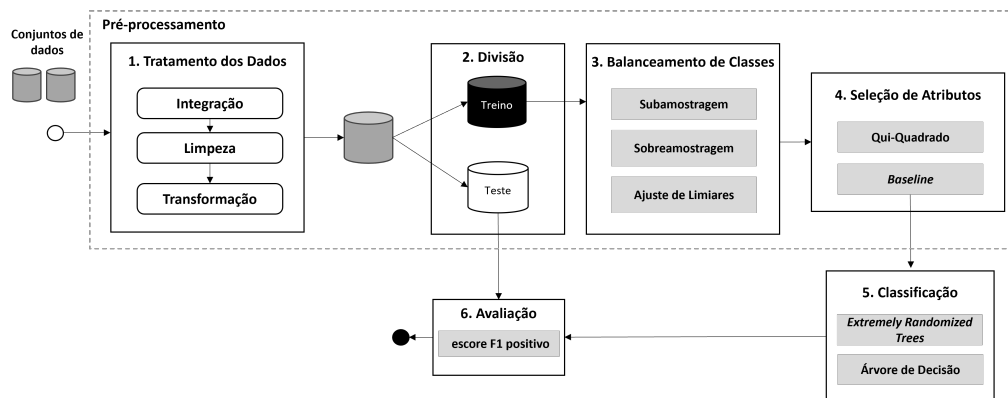


Figura 1. Visão geral da metodologia adotada.

Foram utilizados dois conjuntos de dados públicos desbalanceados de transações de cartão de crédito. O primeiro<sup>3</sup> foi desenvolvido pelo *Institute of Electrical and Electronics Engineers - Computational Intelligence Society* (IEEE-CIS) em parceria com a Vesta Corporation, líder global em serviços de pagamentos. Ele apresenta cerca de 590 mil entradas (em que 3,5% das transações são fraudulentas), com 433 atributos, como valor, data e horário e produto de cada transação, além de e-mail e endereço do comprador. Já o segundo conjunto de dados<sup>4</sup> foi desenvolvido pela empresa *Worldline* e o *Machine Learning Group* (MLG) da *Université Libre de Bruxelles* (ULB) e serviu como conjunto de

<sup>3</sup><https://www.kaggle.com/competitions/ieee-fraud-detection/data>

<sup>4</sup><https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

validação do *pipeline* desenvolvido com base nos dados do IEEE-CIS. Ele apresenta cerca de 285 mil entradas (em que 0,2% das transações são fraudulentas) com 30 atributos.

A fase de pré-processamento inclui as etapas de tratamento de dados, divisão do conjunto de dados em treino e teste utilizando estratificação, balanceamento das classes (transação legítima e fraudulenta) e seleção de atributos. O tratamento de dados inicia-se com a preparação dos dados, integrando os conjuntos de transações e identidades dos compradores para obter um conjunto de dados unificado e completo. Em seguida, remove-se as colunas com mais de 75% dos valores faltantes, além da remoção das colunas com mais do que 75% dos valores únicos, sendo o conjunto IEEE-CIS reduzido de 433 atributos para 91. Por fim, é realizada a transformação dos dados, substituindo valores faltantes de atributos numéricos pela média e atributos categóricos pela moda, e convertendo atributos categóricos em atributos binários utilizando *one-hot encoding*. Ao final dessa última etapa, o conjunto de dados IEEE-CIS ficou com 163 atributos. Finalmente, os dados são divididos em conjuntos de treino e teste utilizando estratificação, garantindo a manutenção da proporção original das classes, essencial para lidar com o desbalanceamento entre transações legítimas e fraudulentas.

A etapa seguinte do pré-processamento consiste em aplicar técnicas para o balanceamento das classes. Foram utilizadas (i) técnicas de superamostragem: *Resample*, ROS, SMOTE, BSMOTE e ADASYN, para aumentar exemplos da classe minoritária e, assim, garantir maior representatividade e diversidade dos dados; (ii) técnicas de subamostragem: RUS, *Near Miss* e IHT, para reduzir exemplos da classe majoritária, focando nos casos mais desafiadores e representativos; e (iii) ajuste de limiares durante o treinamento do modelo para busca de valor ótimo de classificação entre transações genuínas e fraudulentas, que maximize o desempenho desse processo.

Por fim, a última etapa do pré-processamento realiza a seleção de atributos mais relevantes para a detecção de fraudes, tanto para os conjuntos de dados sobreamostrados quanto para os subamostrados. Para isso, utilizou-se a técnica Qui-Quadrado [Laborda and Ryoo, 2021], em que foram selecionados os 100 atributos de maior valor pela técnica no conjunto de dados IEEE-CIS e, no caso, do Worldline-MLG, os primeiros 20 atributos. Para fins de comparação, também foi considerada a abordagem sem seleção de atributos (*baseline*).

Para etapa da classificação, foram utilizados os classificadores ET e AD, pois, ao longo dos experimentos, foram os que demonstraram melhor desempenho em relação a outros previamente selecionados, como RF, CB, GB, AB, RL, KNN e NB. A classificação foi realizada utilizando validação cruzada com estratificação do conjunto de dados. Finalmente, os modelos foram avaliados por meio da métrica escore  $F_1$ .

## 4. Resultados

Os experimentos foram realizados em um computador com processador Intel(R) Core(TM) i5-10400 CPU de 2,90GHz e 130 GB de memória RAM sob o sistema Ubuntu v20.04, e implementados<sup>5</sup> em Python usando as bibliotecas *pandas*<sup>6</sup>, *scikit-*

<sup>5</sup>O código-fonte está disponível em [Anonimizado].

<sup>6</sup><https://pandas.pydata.org/pandas-docs/version/2.0.2/index.html>

*learn*<sup>7</sup>, *numpy*<sup>8</sup> e *imbalanced-learn*<sup>9</sup>. Em ambos os classificadores, ET e AD, foi utilizada a validação cruzada *k-fold* com 10 *folds* e estratificação. No classificador ET, foram utilizadas 100 árvores como estimadores. Os demais parâmetros dos classificadores e técnicas de balanceamento foram os valores padrões das funções nas bibliotecas. A base de dados foi dividida em 80% para treinamento e 20% para teste com estratificação durante a separação. Finalmente, em termos de métricas de desempenho, foi utilizado o escore  $F_1$  (escore  $F_1 = \frac{2*VP}{2*VP+FN+FP}$ ), que se apresenta em função dos Verdadeiros Positivos (VP), Falsos Positivos (FP) e Falsos Negativos (FN), para avaliar a eficácia dos modelos, considerando a natureza desbalanceada dos dados [Hilal et al., 2022; Bhattacharyya et al., 2011].

Inicialmente, foi realizada uma análise exploratória do conjunto de dados da IEEE-CIS para ter conhecimento, por exemplo, da distribuição das médias e modas dos atributos, e avaliar a necessidade de tratamento dos dados. Em seguida, foram feitos experimentos com o intuito de encontrar combinações de técnicas de seleção de atributos, técnicas de balanceamento e modelos de classificação que maximizassem o escore  $F_1$ . Em termos de seleção de atributos, a técnica QQ com filtro dos primeiros 100 atributos e o *baseline* (todos os atributos) demonstraram um desempenho melhor. Já em relação aos classificadores, tanto AD quanto ET demonstraram os melhores desempenhos, com limiares de classificação em torno dos valores 0,25 e 0,55 superiores aos outros limiares, que variaram entre 0,05 e 0,75 na fase de ajustes dos experimentos. Finalmente, em relação às técnicas de balanceamento, os métodos de subamostragem demonstraram desempenho pior que os de superamostragem, mas foram mantidas até a versão final do *pipeline* de dados justamente por conta da proposta desse trabalho de aprofundar o entendimento dessas técnicas. Assim, a versão final do *pipeline*, bem como a Tabela 1, apresentam apenas a combinação de métodos de seleção de atributos (QQ e *baseline*), ajustes de limiares (0,25 e 0,55), técnicas de superamostragem, técnicas de subamostragem e classificadores (AD e ET) com melhores desempenhos. Também são apresentados os resultados em que nenhuma técnica de balanceamento é implementado (N/A). A Tabela 1 mostra resultados tanto do conjunto IEEE-CIS quanto do conjunto Worldline-MLG. Vale ressaltar que, para o conjunto IEEE-CIS, as técnicas de superamostragem chegaram a uma proporção de 50/50 (570 mil / 570 mil) no balanceamento, enquanto as técnicas de subamostragem chegaram a uma proporção de 90/10 (230 mil / 20 mil). O ajuste de limiares não gerou balanceamento, ficando com a proporção de 570 mil transações genuínas e 20 mil fraudulentas. Também importante ressaltar que cada variação da tabela foi executada dez vezes para análise da média e desvio-padrão.

Para o conjunto IEEE-CIS, as combinações que utilizam o ajuste de limiares com ET obtiveram os melhores desempenhos, com a combinação sem seleção de atributos (*baseline*) apresentando o melhor escore  $F_1$  médio, de 0,732. As técnicas de sobreamostragem SMOTE e ADASYN com ET também apresentaram escore  $F_1$  próximos aos melhores obtidos com o ajuste de limiares. Em contrapartida, a técnica de subamostragem IHT obteve sempre um desempenho inferior. No caso do conjunto Worldline-MLG, a combinação que utilizou o ajuste de limiares com ET e sem seleção de atributos (*baseline*) também apresentou os melhores desempenhos para o escore  $F_1$  médio, de 0,862. No entanto, a

<sup>7</sup>[https://scikit-learn.org/stable/whats\\_new/v1.2.html](https://scikit-learn.org/stable/whats_new/v1.2.html)

<sup>8</sup><https://numpy.org/doc/stable/release/1.24.0-notes.html>

<sup>9</sup><https://imbalanced-learn.org/stable/>

Conjunto de dados	Seleção de atributos	Classificador	Balanceamento de classes	escore $F_1$
IEEE-CIS	QQ (top 100)	ET	N/A	0,651 ± 0,002
			Ajuste de limiares (Limiar: 0,30)	<b>0,729</b> ± 0,001
			IHT	0,405 ± 0,001
			SMOTE	0,685 ± 0,003
	AD	N/A	0,542 ± 0,002	
		Ajuste de limiares (Limiar: 0,55)	0,542 ± 0,002	
		IHT	0,291 ± 0,001	
		ROS	0,542 ± 0,002	
baseline (163 atributos)	ET	N/A	0,549 ± 0,002	
		Ajuste de limiares (Limiar: 0,30)	<b>0,732</b> ± 0,002	
		IHT	0,423 ± 0,002	
		AdaSyn	0,681 ± 0,002	
AD	N/A	0,541 ± 0,004		
	Ajuste de limiares (Limiar: 0,55)	0,549 ± 0,002		
	IHT	0,293 ± 0,001		
	ROS	0,539 ± 0,001		
Worldline-MLG	QQ (top 20)	ET	N/A	0,830 ± 0,005
			Ajuste de limiares (Limiar: 0,40)	0,839 ± 0,009
			IHT	0,779 ± 0,009
			BSMOTE	<b>0,860</b> ± 0,008
	AD	N/A	0,690 ± 0,022	
		Ajuste de limiares (Limiar: 0,55)	0,695 ± 0,014	
		IHT	0,556 ± 0,014	
		ROS	0,696 ± 0,022	
baseline (30 atributos)	ET	N/A	0,700 ± 0,021	
		Ajuste de limiares (Limiar: 0,25)	<b>0,862</b> ± 0,006	
		IHT	0,760 ± 0,006	
		SMOTE	0,843 ± 0,008	
AD	N/A	0,695 ± 0,018		
	Ajuste de limiares (Limiar: 0,55)	0,710 ± 0,016		
	IHT	0,559 ± 0,011		
	ROS	0,723 ± 0,012		

**Tabela 1. Resultados dos experimentos**

técnica de superamostragem BSMOTE com ET e seleção dos 20 melhores atributos pelo QQ também alcançou resultado semelhante, de 0,860. A abordagem com o IHT também apresentou, em média, um desempenho inferior.

### 5. Conclusões

Este artigo apresentou a aplicação de técnicas de balanceamento em dois conjuntos diferentes de bases de dados, com métricas de desempenho apropriadas para dados desbalanceados, com são os de transações fraudulentas. A metodologia proposta foi capaz de mostrar, comparativamente, os efeitos de sua aplicação. Os ajustes de limiares na faixa de 0,25 a 0,55 demonstraram desempenho superior às técnicas de subamostragem e superamostragem, embora a própria superamostragem tenha tido escore  $F_1$  relevante em alguns casos. Nesse caso, SMOTE e ADASYN demonstraram melhor escore, com as técnicas de subamostragem revelando-se piores nesse sentido. Como trabalhos futuros, pode-se explorar a abordagem híbrida de ajustes de limiares em conjunto com as técnicas de balanceamento, explorar os próprios classificadores por meio da otimização de seus hiper-parâmetros, e propor melhorias nas técnicas de balanceamento avaliadas - ou, até mesmo, novos métodos para tal.

### Referências

Amit Singh, R. K. R. and Tiwari, A. (2022). Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(4):571–598.

- Bhagwani, H., Agarwal, S., Kodipalli, A., and Martis, R. J. (2021). Targeting class imbalance problem using gan. In *5th Inter. Conf. on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pages 318–322.
- Bhattacharyya, S. et al. (2011). Data mining for credit card fraud: A comparative study. *Decis. Support Syst.*, 50:602–613.
- Carcillo, F. et al. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557:317–331.
- Ghaleb, F. A. et al. (2023). Ensemble synthesized minority oversampling-based generative adversarial networks and random forest algorithm for credit card fraud detection. *IEEE Access*, 11:89694–89710.
- Gupta, P. et al. (2023). Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques. *Procedia Computer Science*, 218:2575–2584. International Conference on Machine Learning and Data Engineering.
- Hasib, K. M. et al. (2020). A survey of methods for managing the classification and solution of data imbalance problem. *Journal of Computer Science*, 16(11):1546–1557.
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.
- Hilal, W. et al. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 193:116429.
- Ileberi, E. et al. (2021). Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost. *IEEE Access*, 9:165286–165294.
- Jahnvi, Y. et al. (2023). A novel ensemble stacking classification of genetic variations using machine learning algorithms. *International Journal of Image and Graphics*, 23.
- Karthika, J. and Senthilselvi, A. (2023). An integration of deep learning model with navo minority over-sampling technique to detect the frauds in credit cards. *Multimedia Tools Appl.*, 82(14):21757–21774.
- Laborda, J. and Ryoo, S. (2021). Feature selection in a credit scoring model. *Mathematics*, 9(7).
- Leevy, J., Johnson, J., Hancock, J., and Khoshgoftaar, T. (2023). Threshold optimization and random undersampling for imbalanced credit card data. *Journal of Big Data*, 10.
- Makki, S. et al. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7:93010–93022.
- Muaz, A. et al. (2020). A comparison of data sampling techniques for credit card fraud detection. *International Journal of Advanced Computer Science and Applications*, 11.
- Prabha, D. P. and Priscilla, C. V. (2024). Estimation of optimal threshold shifting to handle class imbalance in credit card fraud detection using machine learning techniques. In *American Institute of Physics Conference Series*, volume 2802, page 120014. AIP.
- Priscilla, C. V. and Prabha, D. P. (2020). Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. In *3rd Inter. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, page 1309–1315.
- Sisodia, D. S., Reddy, N. K., and Bhandari, S. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 2747–2752.
- Sun, Y. et al. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719.
- Xie, Y., Li, A., Gao, L., and Liu, Z. (2021). A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wireless Communications and Mobile Computing*, 2021(1):2531210.
- Zhang, F. et al. (2019). Gmm-based undersampling and its application for credit card fraud detection. In *International Joint Conference on Neural Networks*, pages 1–8.