Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241485

Transaction fraud detection by CatBoost model with feature engineering

Jiahan Shi

School of computer science and technology, Harbin Institute of Technology, Shenzhen, China

shijiahan@stu.hit.edu.cn

Abstract. In the digital era, the ubiquitous nature of online transactions has placed a spotlight on the imperative for robust fraud detection systems. This research tackles this pressing issue, with a particular emphasis on the nuanced process of feature engineering tailored for the IEEE-CIS dataset, a representative sample of contemporary transactional behaviors. The enhancement of data attributes, through meticulous feature engineering, acts as the bedrock for the application of the CatBoost model - a gradient-boosting technique revered for its precision. The paper dedicates significant attention to both these pivotal stages, showcasing the synergistic effect they have when applied in tandem. With this refined data and sophisticated modeling, the proposed method manifests exceptional performance, establishing a new benchmark when juxtaposed with other gradient-boosting methodologies. Conclusively, this study offers valuable insights for enhancing online transaction security and sets the stage for further innovation in fraud detection within the machine learning community.

Keywords: Transaction Fraud Detection, CatBoost, Feature Engineering.

1. Introduction

In the contemporary digital era, the rise in online transactions significantly amplifies the risks of financial fraud. As the volume of digital transactions grows, malicious actors deploy numerous tactics to exploit vulnerabilities targeting transaction systems, ranging from identity theft to conducting unauthorized transactions [1]. With each fraudulent transaction posing potential massive losses, the imperative for efficient transaction fraud detection has significantly risen for organizations and their customers [2]. This ensures the safety and dependability of every transaction made. However, the challenge intensifies when trying to distinguish authentic transactions from fraudulent ones, particularly as deceitful strategies aimed at transactions continually evolve.

In recent years, there has been significant academic attention and progress in the field of transaction fraud detection. Historically, conventional approaches often depended on rule-based systems, whereby pre-established criteria were used to identify potentially dubious transactions. Although these systems have shown some effectiveness, they often encounter difficulties responding to emerging fraud strategies, resulting in elevated percentages of false positives [3, 4]. Machine learning, characterized by its adaptable and data-centric characteristics, has fundamentally transformed transaction fraud detection [5-8]. Multiple research projects have been utilized to explore innovative fraud detection methodologies. Random Forests, known for their ensemble learning mechanism and

DOI: 10.54254/2755-2721/48/20241485

ability to handle large feature sets, have seen significant application [9]. Shaohui et al. employed a Random Forest classifier, highlighting its effectiveness in capturing non-linear transaction patterns while reducing overfitting [10]. Neural Networks have been gaining traction in the fraud detection domain. Several studies used deep neural network models, which yielded remarkable levels of accuracy [8, 11, 12]. These authors' research showcased the model's capacity to effectively identify intricate and multi-dimensional data patterns inherent in activities related to fraud detection. Gradient-boosting frameworks have been another focal point. XGBoost and LightGBM have found favor due to their ability to handle imbalanced data and iterative learning mechanisms [13-16]. Recently, CatBoost has emerged as a promising candidate [17]. The work conducted by Chen and Han demonstrated the effectiveness of utilizing CatBoost, highlighting its capacity to provide reliable outcomes [18]. Although its performance is commendable, there is a noticeable scarcity of comprehensive research delving into its optimization and potential applications in various domains.

This research provides an approach that combines advanced feature engineering approaches with a CatBoost-based methodology. By combining the benefits of gradient-boosting frameworks with the subtleties of customized data preparation, this method seeks to provide a holistic response to the problems associated with fraud detection.

2. Dataset Description

In the realm of digital transactions, the rise in fraudulent activities has necessitated the development of advanced fraud detection mechanisms. Leveraging machine learning and artificial intelligence, these detection systems have emerged as pivotal tools for ensuring the security of online financial transactions. The efficacy of fraud detection mechanisms is contingent upon the richness and thoroughness of the training data utilized, and the IEEE Computational Intelligence Society, in collaboration with Vesta, has developed the IEEE-CIS dataset to meet these requirements. The dataset provides insight into the complex and subtle features of current trading practices.

This dataset comprises four distinct tables, named begin with train and test respectively. This research primarily examines the train tables due to the lack of labels in the test tables, subsequently omitting the prefix train for simplicity.

Transaction table. The transaction table is a comprehensive source of information within the IEEE-CIS dataset, encompassing 394 columns. Detailed specifications and descriptions of each of these columns can be found in Table 1.

Table 1. Detailed descriptions of the transaction table.

Column	Column Description
TransactionID	ID of the given transaction
isFraud	Binary classification target
TransactionDT	Time delta
TransactionAMT	Payment amount (USD)
ProductCD	Product code
card1 to card6	Payment card information
P_emaildomain	Email domain of purchaser
R_emaildomain	Email domain of Receiver
addr1 and addr2	Address
dist1 and dist2	Distance
C1 to C14	Anonymous features
D1 to D15	Time delta
M1 to M9	Anonymous features
V1 to V339	Engineered features

Columns C1 to C14 refer to counting information, represent counting information, with their specific meanings masked for confidentiality or data sensitivity. Columns M1 to M9 contain match-related data, indicating congruence between transactional details such as cardholder names and

DOI: 10.54254/2755-2721/48/20241485

addresses [19]. Columns V1 to V339, crafted by the original dataset providers, are engineered features derive from advanced data processing and transformation techniques. The presence of these columns indicates the depth of preprocessing and feature extraction invested into the dataset, making it suitable for high-quality fraud detection models.

Identity table. The identity table is a crucial element inside the IEEE-CIS dataset, including important indicators for identity verification. This table has a total of 41 columns and focuses on identifying information. It covers facts related to network connections such as IP addresses, Internet Service Providers (ISPs), and Proxy information. It also incorporates digital signatures comprising user-agent data, browsers, operating systems, and their corresponding versions [19]. A comprehensive description of each column's specification can be found in Table 2.

Table 2. Detailed descriptions of the identity table.

Column Name	Description
TransactionID	ID of the given transaction
id_01 to id_38	Anonymous identity related information
DeviceType	Device type information, such as mobile and desktop
DeviceInfo	Detailed information of device

3. Methodology

The success of a fraud detection system depends not only on sophisticated models but also on the careful choice of input features. This section details the feature engineering techniques that enhance the quality and relevance of the dataset's attributes, setting a robust foundation for subsequent modeling. Subsequently, the CatBoost model is introduced, focusing on comparing it with other gradient-boosting models and discussing the specific parameters and configurations tailored for the task. Both these components, in tandem, contribute to the method's enhanced ability to discern fraudulent patterns within the dataset.

3.1. Feature Engineering

Before utilizing machine learning techniques, feature engineering plays a crucial role in converting raw data into a structured manner that enhances the effectiveness of prediction models. When faced with a vast dataset such as IEEE-CIS, it is essential to shape this data to fully use its predictive capabilities systematically. Two primary strategies underpin the feature engineering phase. First, given the profusion of features, it is essential to distill the data, ensuring only the most informative attributes are retained, and redundant or less pertinent ones are eschewed. Second, while traditional preprocessing methods, advocate for removing outliers, this method diverges by intentionally preserving them. Recognizing that outliers can often signify genuine anomalies in fraud detection, their inclusion becomes paramount in shaping a model attuned to detecting truly anomalous transactions. The coalescence of these strategies is geared towards refining the dataset into a more potent tool for fraud detection.

Efficient feature engineering is paramount for extracting meaningful insights from the transaction table. Addressing the temporal dimensions of the data is a crucial aspect of this. Columns D1 to D15, representing time-related deltas such as the interval since the previous transaction, were meticulously normalized based on the time delta to ensure that any potential distortions or variations arising from different time increments are neutralized. By doing so, these columns provide a consistent and relative understanding of transaction frequency and timing, which is vital for detecting anomalies or suspicious patterns over time. In addition to handling temporal data, the categorical aspect was also addressed. Columns ProductCD, card4, and card6, given their fixed cardinalities, underwent one-hot encoding, transforming them into a format more amenable to analytical computations.

A crucial challenge was the vast number of engineered columns V1 to V339. Insights from a strategy from a competition winner's Kaggle kernelilluminate the path to tackle this [20]. Notably, many groups of V columns exhibit strong inter-correlation. By conducting a correlation analysis and

DOI: 10.54254/2755-2721/48/20241485

setting a coefficient threshold at 0.75, it becomes possible to identify representative subsets. For instance, the correlation matrix for V1 to V11, displayed in Figure 1, demonstrates that the subset {V1, V3, V4, V6, V8, V11} can represent the main features of the entire group. This approach effectively reduces the number of V columns from 339 to 124, improving computing efficiency. Other categorical features outside the abovementioned columns are transformed through label encoding, creating a standardized format. Lastly, missing values are filled by a placeholder value of -999, which is not a cosmetic fix but intended to expedite computational processes.

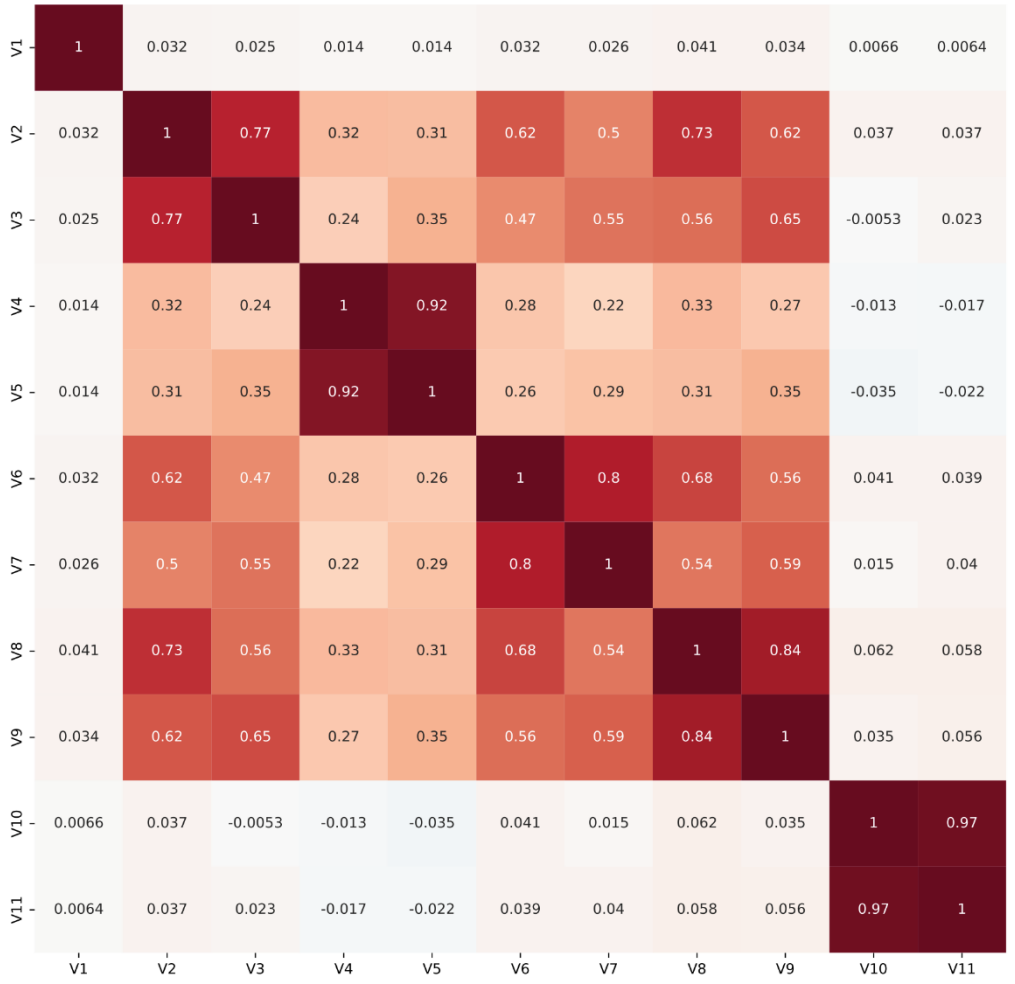

Figure 1. Correlation matrix

Within the identity table, most columns ranging from id_01 to id_38 have been label encoded for standardization. Exceptions are made for device-related columns such as id_23, id_30, and id_34. These specific columns are amalgamated with DeviceType and DeviceInfo columns, deriving new features such as device_name and device_version. Consistent with the approach for the transaction table, missing values in this table are filled with -999.

3.2. CatBoost-based Model

The CatBoost algorithm, derived from "Category Boosting", has garnered recognition for its high performance in handling categorical data, making it a compelling choice for fraud detection tasks that often deal with a mix of categorical and numerical features.

The inherent ability of CatBoost to naturally handle categorical data, coupled with its resistance to overfitting due to built-in regularization, makes it particularly adept for the fraud detection scenario,

DOI: 10.54254/2755-2721/48/20241485

where feature relationships can be intricate and non-linear. Among various gradient boosting algorithms available, CatBoost is selected for this research over other popular models such as XGBoost and LightGBM. The rationale behind this selection is multifaceted:

Regularization: CatBoost incorporates built-in support for L2 regularization, which can help prevent overfitting, especially when dealing with datasets having numerous features. This contributes to more generalizable models, which is essential for fraud detection where novel fraudulent patterns may emerge over time.

Robustness to Noisy Data: Financial transaction data, like the IEEE-CIS dataset, can sometimes be noisy for reasons such as manual data entry errors, system glitches, or discrepancies in transaction recording. CatBoost's algorithm is structured to be less sensitive to noise in the data, ensuring that the model remains robust in its predictions.

Comparative Baseline Assessment: To validate the efficiency of CatBoost, comparative experiments were conducted using baseline models developed with XGBoost and LightGBM. Both these models are well-regarded in the machine learning community and have demonstrated their capabilities in various applications. However, when applied to the IEEE-CIS dataset, it was observed that CatBoost outperformed or matched these models in key performance metrics [18].

While both XGBoost and LightGBM have their own merits and have shown promising results in various applications, the above-mentioned qualities made CatBoost the preferred choice for this specific research context.

To obtain optimal performance from the CatBoost model, meticulous parameter tuning was conducted. The specific parameters employed in this research are detailed in Table 3; parameters such as iterations, learning_rate, and depth significantly impact the behavior of the model and thus affect its predictive capacity.

Table 3. Fraud detection CatBoost model parameters.

Parameter Name	Description	Parameter Value
iterations	Number of boosting iterations	10000
learning_rate	Model's learning rate	0.03
depth	Depth of the trees	11
loss_function	Objective function	Logloss
random_seed	Seed for random number generator	42
eval_metric	Metric for model's evaluation	AUC
metric_period	Iteration logging frequency	500
od_wait	Number of iterations for early stopping	500
task_type	Hardware preference for task	GPU

4. Experiments and Results

This section delves into a detailed evaluation of the proposed CatBoost-based method utilizing the IEEE-CIS dataset. Considering rising digital transaction frauds, establishing a robust detection system is paramount. The main goal of this study is to assess the efficacy of the proposed method in identifying suspicious transaction patterns through the process of training and, thereafter, comparing its performance with other widely used models. This study also evaluated the computational performance, precision, and dependability of the suggested method relative to other prevalent machine learning models using the IEEE-CIS dataset. The experiments were conducted using a Ubuntu 22.04 Linux machine that was equipped with an NVIDIA 3090 GPU.

In this study, the performance metrics employed encompassed the Area Under the Receiver Operating Characteristic curve (AUC-ROC Score) and Accuracy. For fraud detection, the AUC-ROC metric is of paramount importance. The AUC-ROC score, commonly employed in binary classification tasks, offers a holistic insight into a model's performance across varied classification thresholds. In the realm of digital transactions, fraudulent activities, while infrequent, pose a significant threat to the integrity and trust of the system. As such, it becomes imperative for detection

DOI: 10.54254/2755-2721/48/20241485

models to have a high true positive rate, ensuring that most fraudulent transactions are caught. At the same time, it's equally crucial to ensure that the false positive rate remains low to avoid mistakenly flagging legitimate transactions as fraudulent, which could erode user trust and lead to unnecessary investigations.

To improve binary classification performance on the IEEE-CIS dataset, this research compares the effectiveness of various models such as Logistic Regression, SVM, XGBoost, LightGBM, and the baseline CatBoost model, with the method proposed in this paper. A detailed comparison of their outcomes can be found in Table 4.

Table 4. Performance analysis: proposed technique versus other models on the IEEE-CIS dataset.

Model Name	AUC-ROC Score	Accuracy
Logistic Regression	0.862	0.931
SVM	0.907	0.958
XGBoost	0.952	0.981
LightGBM	0.955	0.982
CatBoost Baseline	0.971	0.983
CatBoost with Feature Engineering	0.974	0.988

Upon analyzing the metrics of various models on the IEEE-CIS dataset, several observations emerge. While Logistic Regression and SVM offer respectable AUC- ROC scores of 0.862 and 0.907, respectively, their performance is visibly overshadowed by gradient boosting methods like XGBoost, LightGBM, and CatBoost. Among these, the CatBoost baseline model achieves an impressive AUC-ROC score of 0.971. Notably, when enhanced with feature engineering, the CatBoost model outperforms all other models, attaining 0.974 of AUC-ROC score and 0.988 of accuracy rate. These results underline the significance of advanced model optimization and the pivotal role of feature engineering in achieving superior fraud detection capabilities.

Feature importance analysis pinpoints the key variables influencing the CatBoost model's decisions. The 20 most influential features are depicted in Figure 2, which offers a clear understanding of the variables that the CatBoost model deemed most influential in the fraud detection task:

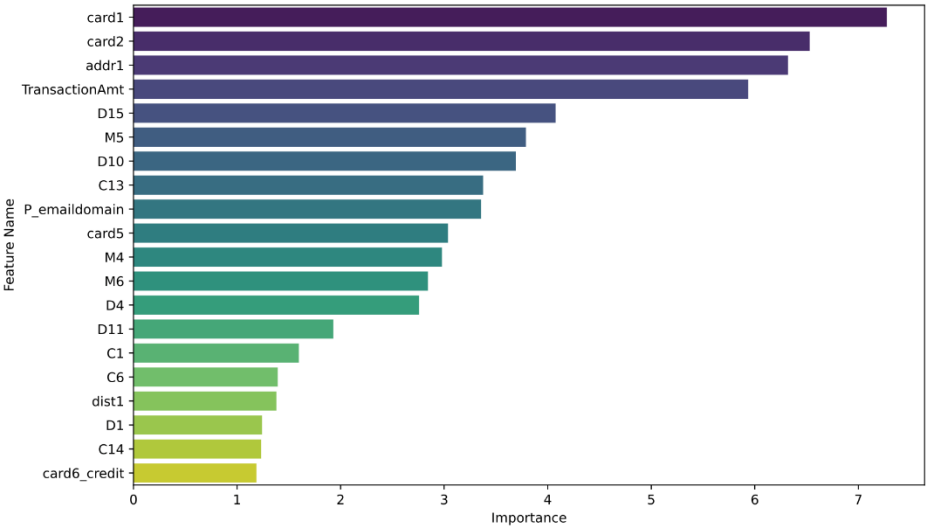

Figure 2. Top 20 features ranked by importance

Card-Related Features: Two of the most import features are card1 and card2, which suggest that details associated with the payment card are critical predictors when identifying potentially fraudulent transactions.

DOI: 10.54254/2755-2721/48/20241485

Address and Transaction Amount: Both addr1 and TransactionAmt rank high in importance. The address might relate to transaction patterns across different geographic regions, while the transaction amount could indicate patterns typical of fraudulent activity.

Email Domain: The P_emaildomain feature shows that the email domain of the purchaser might be an indicative factor, with certain domains possibly being more prevalent in suspicious activities.

Time Delta & Counting Features: Features like D15, D10, C13, and D4 highlight the significance of time between transactions and specific counting metrics, hinting at transaction frequencies and intervals.

The model's emphasis on a mix of card details, transactional behavior, user details, and device information underlines the multifaceted nature of fraud detection and the comprehensive approach required to predict such activities.

5. Conclusions

This study illuminated the challenges and complexities involved in fraud detection. By employing advanced CatBoost model and rigorous feature engineering, the research established the critical role that specific features play in determining transactional legitimacy. The prominence of features like card details, address, transaction amounts, and device information underscored the multi- dimensional aspect of fraud detection. Additionally, when juxtaposing the efficacy of different models, it became evident that the CatBoost model, augmented with meticulous feature engineering, stood out as an exceptionally proficient instrument for tackling this challenge. As online transactions continue to evolve, so will the tactics of fraudsters, making continual research and model advancement in this domain an ongoing necessity.

Looking ahead, there are several avenues of exploration that could further enhance the efficiency of fraud detection systems. One promising direction is the incorporation of deep learning architectures, which might uncover more intricate patterns within transaction data. Additionally, the potential of unsupervised and semi-supervised learning techniques, especially in scenarios with limited labeled data, warrants investigation. Lastly, real-time fraud detection, where swift decision-making is crucial, offers a challenging yet rewarding frontier for future endeavors in this domain.

References

[1] Dan Sun, Wei-li Shi, Lan-xiang Rao, Sha-sha Meng, Xiao-ming Guo, and Yi-lun Li. Credit Card Fraud Detection Method Based on Improved SMOTE+ENN and XGBoost Algorithm. Computer and Modernization, 0(09):111–118, 2022.

[2] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. Journal of Network and Computer Applications, 68:90–113, 2016. ISSN 1084-8045.

[3] Zhong Li, Xiaolong Jin, Chuanzhi Zhuang, et al. Overview on graph based anomaly detection. Journal of Software, 32(1):167–193, 2020.

[4] Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, and Yo-Ping Huang. Survey of fraud detection techniques. In IEEE International Conference on Networking, Sensing and Control, 2004, volume 2, pages 749–754. IEEE, 2004.

[5] Qinxin Chen. Machine Learning Methods in Credit Card Fraud Detection: A Comparative Study. China High-Tech, 24:5, 2018.

[6] Xiangrong Shi, Pengsai Guo, Qi Zheng, et al. Application of Ensemble Learning in Consumer Finance Auditing: A Case Study of Credit Card Fraud Detection Using Random Forest. Commercial Accounting, 2022(15):6.

[7] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. Random forest for credit card fraud detection. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), pages 1–6, 2018.

[8] AI Sergadeeva, DS Lavrova, and DP Zegzhda. Bank fraud detection with graph neural networks. Automatic Control and Computer Sciences, 56(8):865–873, 2022.

[9] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

DOI: 10.54254/2755-2721/48/20241485

[10] Du Shaohui, GuanWen Qiu, Huafeng Mai, and Hongjun Yu. Customer transaction fraud detection using random forest. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pages 144–147, 2021.
[11] Lei Zhang, KaiFeng Ma, Fang Yuan, and WenJun Fang. A tabnet based card fraud detetion algorithm with feature engineering. In 2022 2nd International Conference on Consumer Elec- tronics and Computer Engineering (ICCECE), pages 911–914, 2022.
[12] Xiong Kewei, Binhui Peng, Yang Jiang, and Tiying Lu. A hybrid deep learning model for online fraud detection. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pages 431–434, 2021.
[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.
[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.
[15] Yixuan Zhang, Jialiang Tong, Ziyi Wang, and Fengqiang Gao. Customer transaction fraud detection using xgboost model. In 2020 International Conference on Computer Engineering and Application (ICCEA), pages 554–558, 2020.
[16] Dingling Ge, Jianyang Gu, Shunyu Chang, and JingHui Cai. Credit card fraud detection using lightgbm model. In 2020 International Conference on E-Commerce and Internet Technology (ECIT), pages 232–236, 2020.
[17] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.
[18] Yeming Chen and Xinyuan Han. Catboost for fraud detection in financial transactions. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pages 176–179, 2021.
[19] IEEE-CIS fraud detection data description (details and discussion). https://www.kaggle.com/competitions/ieee-fraud-detection/discussion/101203, 2019.
[20] Chris Deotte. EDA for columns V and ID. https://www.kaggle.com/code/cdeott e/eda-for-columns-v-and-id/, 2019.