An Empirical Analysis of Machine Learning for Fraud Detection in Diverse Financial Scenarios

Dongqing Lin[1,a,*]

[1]Carey Business School, Johns Hopkins University, Baltimore MD 21202, USA
a. dlin38@jh.edu
**corresponding author*

Abstract: Fraud detection is important in various domains. As fraud patterns become complex, it is critical to develop accurate and efficient detection systems based on machine learning. However, in different scenarios, there are different key points for fraud detection, and the best model and important features of machine learning may also be different. Therefore, this study aims to explore the relationship between the key points in each scenario and the best model and important features of machine learning corresponding to the scenario. This research firstly used scikit-learn to perform machine learning on the datasets of three different scenarios, online payment, credit card, bank account, and find out important features to analyze the applicability of machine learning models in different scenarios and the reasons behind it. Based on the characteristics of each kind of data, the machine learning model and the characteristics of the fraud itself in different scenarios, exploring the explain ability of combining machine learning with human knowledge. The study found that Light GBM is suitable for all scenarios, because it can capture complex relationships in data, and its high efficiency of dealing with imbalanced data. The research identifies distinct important features of each domain combined with human knowledge, provides insight into potentially fraudulent activity. This will provide a certain theoretical basis for developing a more accurate and efficient fraud detection system, improving the performance of human-computer interaction in the system, or applying the human-in-the-loop model.

Keywords: machine learning, fraud detection, online payment, credit card, bank account

1. Introduction

Fraud detection has emerged as a critical task in various domains, including online payment systems, credit card transactions, bank accounts. With the increasing complexity and diversity of fraud patterns, the development of accurate and efficient fraud detection systems has become crucial. Machine learning models have proven to be effective in detecting fraud in different scenarios [1]. These models leverage the power of algorithms and statistical techniques to analyze data and identify fraudulent activities.

There is some research and applications of machine learning in the field of payment and credit card fraud detection. These models leverage large dynamic datasets to track consumer behavior and identify suspicious patterns.

Existing research mainly includes developing algorithms for the performance of machine learning models, comparative studies of the performance of different machine learning algorithms or models in an application scenario, and research on the interpretability of machine learning results in an application scenario. Alarfaj et al. studied algorithms based on machine learning and convolutional neural network architectures to improve fraud detection performance [2]. Kolodiziev et al. studied the use of automated machine learning and big data analysis algorithms to detect fraud in digital payment systems, which is a comprehensive and effective model of fraud [3]. Itri et al. studied comparative performance analysis of machine learning for detecting auto insurance fraud [4]. Awoyemi et al. also conducted similar comparative analysis research on credit cards [5]. Lin et al. studied the interpretability group SHAP of the risk detection model [6]. Current research fails to comprehensively address the generalizability of different machine learning models to detect fraud in multiple scenarios. Furthermore, current research lacks the study of the nuanced characteristics of datasets within each domain. This is because current research does not thoroughly investigate the applicability of these models in various fraud detection scenarios in combination with traditional human knowledge. Understanding the root causes behind fraudulent activity requires a more holistic approach that combines human expertise with machine learning algorithms.

This research is trying to improve fraud detection systems from another point of view. This research combines human knowledge techniques and machine learning techniques for fraud detection in multiple scenarios and correlates them. This study aims to analyze the applicability of different machine learning models in detecting fraud across different scenarios and comparing them. By understanding the characteristics of the dataset within each domain, the reasons behind the selection of specific machine learning models for effective fraud detection can be identified. Additionally, this research explores the interpretability of combining machine learning techniques with human knowledge to gain insights into the underlying reasons behind fraudulent activities in different contexts. This research aims to contribute to the understanding of fraud detection by analyzing the significance of selected models and important features in real-life scenarios.

2. Data

For the three scenarios, online payment, bank account, credit card, the datasets used in this study are: Online Payments Fraud Detection Dataset, Bank Account Fraud Dataset Suite (NeurIPS 2022), Credit Card Transactions Fraud Detection Dataset provided by Kaggle [7].

Each dataset has more than 1,000,000 data and more than 10 features, and their problem type is classification (See Table 1).

Table 1: Basic characteristics of the datasets.

Type of Scenario/Dataset	Number of Instances	Number of Features
Online Payment	6,362,620	11
Bank Account	1,000,000	32
Credit Card	1,296,675	22

The target features are "isFraud", "fraud_bool", "is_fraud"(the fraud label and a value of 1 if the application is classified as fraud and 0 if it is considered legitimate) in each dataset. First counting the number of fraudulent behaviors. For every data set, the number of fraudulent behaviors is far less than the number of normal behaviors. Which means, the question of balance must be considered (See Table 2).

Table 2: Comparisons between fraud case and normal case.

Type of Scenario/Dataset	Fraud Transactions	Normal Transactions
Online Payment	8,213	6,354,407
Bank Account	11,029	988,971
Credit Card	7,506	1,289,169

3. Method

Scikit-learn in python was used throughout the process. These datasets cover various fraud scenarios. For these datasets, the Number of Instances, Features, amount and normal transaction, relationship between variables of the data can be observed. Subsequently, a series of machine learning models, Logistic Regression, Decision Tree, Naive Bayes, k Nearest Neighbors, Random Forest, XGBoost, Light GBM, were applied to evaluate their performance in fraud detection in different domains and feature importance in datasets. Finally, performing data analysis on important features.

In order to integrate human knowledge into the model, feature engineering and domain expertise were used to enhance interpretability and study the significance of the selected model and important features in real-life fraud detection.

4. Result

“Logistic Regression”, “Decision Tree”, "Naive Bayes", "k Nearest Neighbors", "Random Forest", "XG Boost", "Light GBM" classifiers were used for cross-validation on data. Light GBM is the classifier with the highest average accuracy in all scenarios (See Table 3).

Table 3: Average Accuracy for each dataset.

Classifier	Online Payment	Bank Account	Credit Card
Logistic Regression	0.969	0.864	0.944
Decision Tree	0.969	0.839	0.958
Naive Bayes	0.446	0.712	0.977
k Nearest Neighbors	0.983	0.821	0.989
Random Forest	0.969	0.943	0.935
XG Boost	0.986	0.915	0.964
Light GBM	0.998	0.983	0.977

4.1. Online Payment Fraud Detection

The relationship between the two sets of variables, “oldbalanceorg” (balance before the transaction) and “newbalnceorg” (balance after the transaction), and “oldbalancedest” (initial balance of recipient before the transaction) and “newbalancedest” (the new balance of the recipient after the transaction), is highly correlated between two, as shown by the correlation heatmap (See Fig. 1). For variables “isfraud”, “oldbalancedest” and “newbalancedest” are a little more correlated with it.

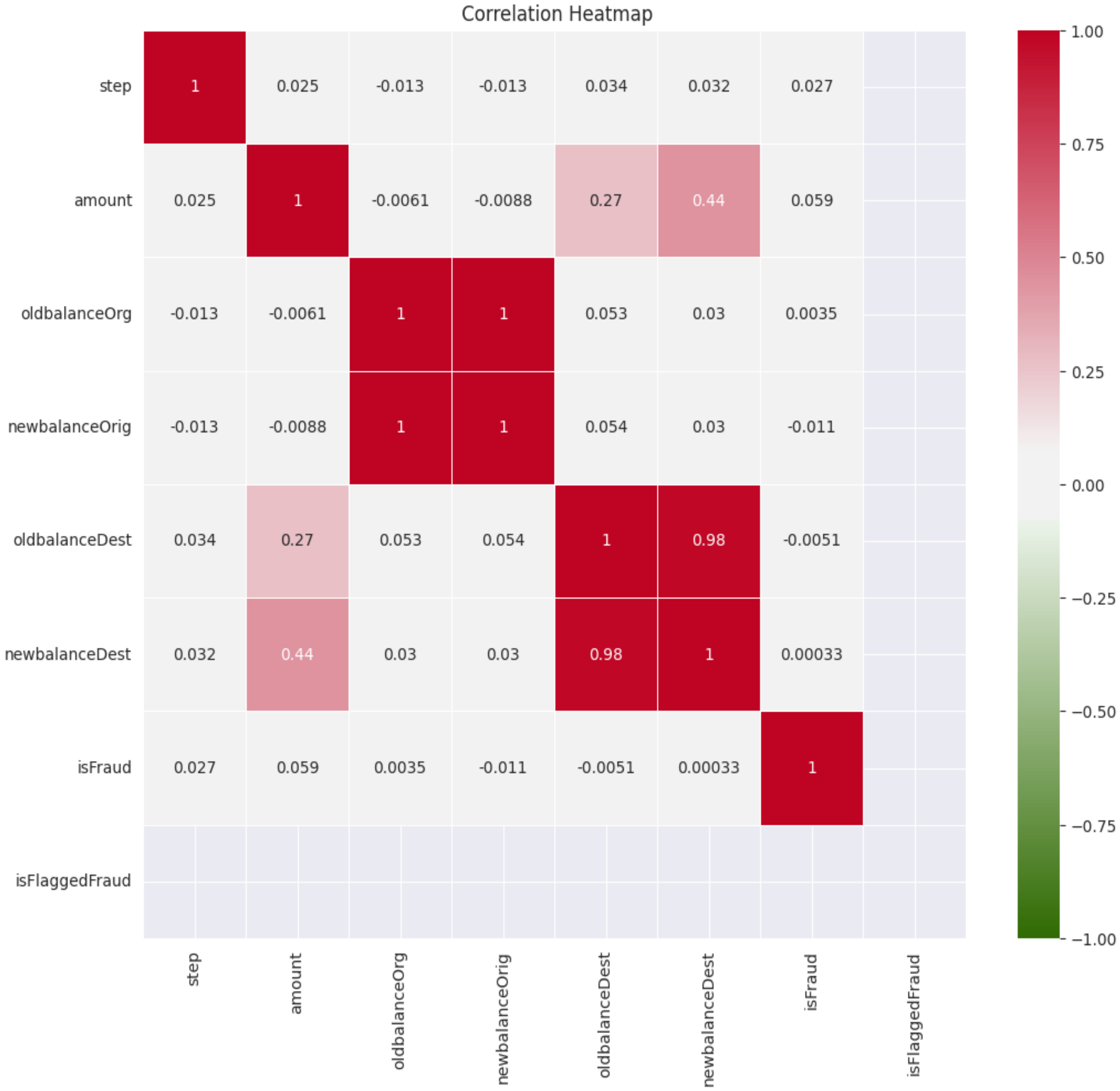


Figure 1: Correlation heatmap of online payment fraud detection.

Scikit-learn was used to calculate and visualize the feature importance for a random forest classifier. Feature importance indicate the relative importance of each feature in predicting the target variable. Visualizing feature importance can help identify the most influential features in the model. As can be seen from the image (See Fig. 2), balance before the transaction, the new balance of the recipient after the transaction, and the amount of the transaction are the three most important features.

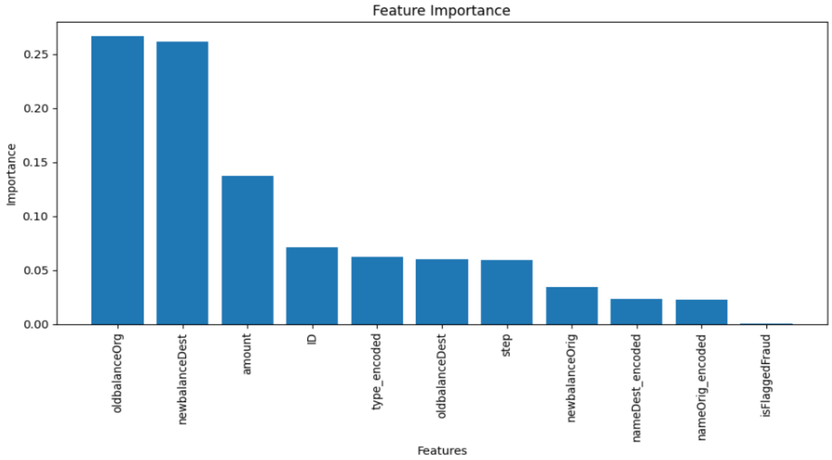


Figure 2: Feature importance of online payment fraud detection.

After specific analysis, fraudulent transactions are more likely to occur when the value of the old balance of origin account is above 30,000 (See Fig. 3); the value of the new balance of the recipient after the transaction is close to 0 (See Fig. 4); or the value of the amount of the transaction is above 30,000 (See Fig. 5).

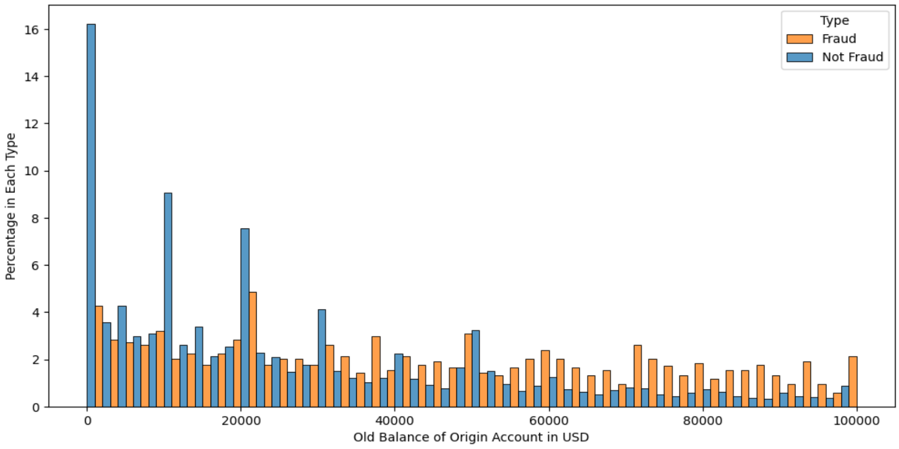


Figure 3: Visualization of old balance of origin account.

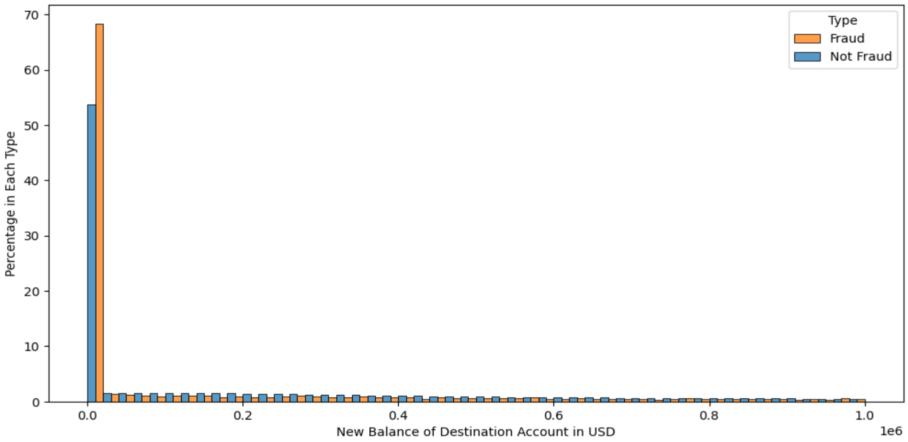


Figure 4: Visualization of new balance of the recipient after the transaction.

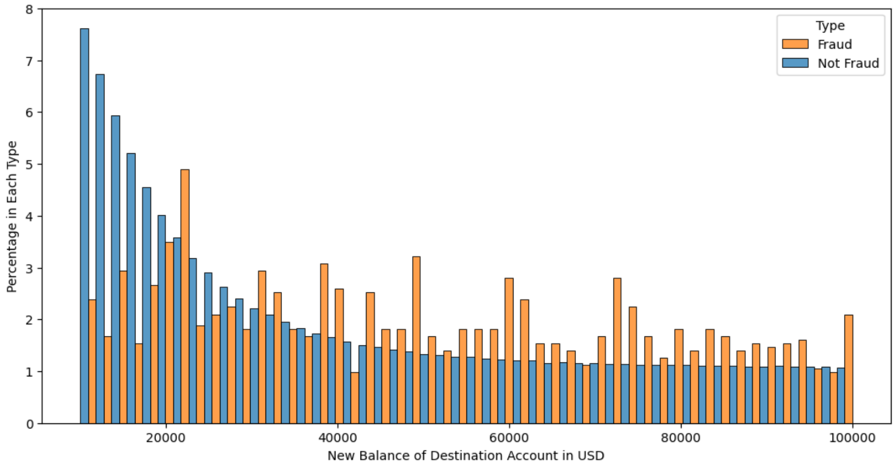


Figure 5: Visualization of value of the amount of the transaction.

4.2. Bank Account Fraud Detection

The relationship between the two sets of variables (See Fig. 6), the four variables velocity_24h, velocity_4w, velocity_6h (Velocity of total applications made in the last 24 hours/4 weeks/24 hours), and zip_count_4w (Number of applications received from the same zip code within the last 4 weeks) have a strong correlation, and they all represent the level of velocity of total applications made.

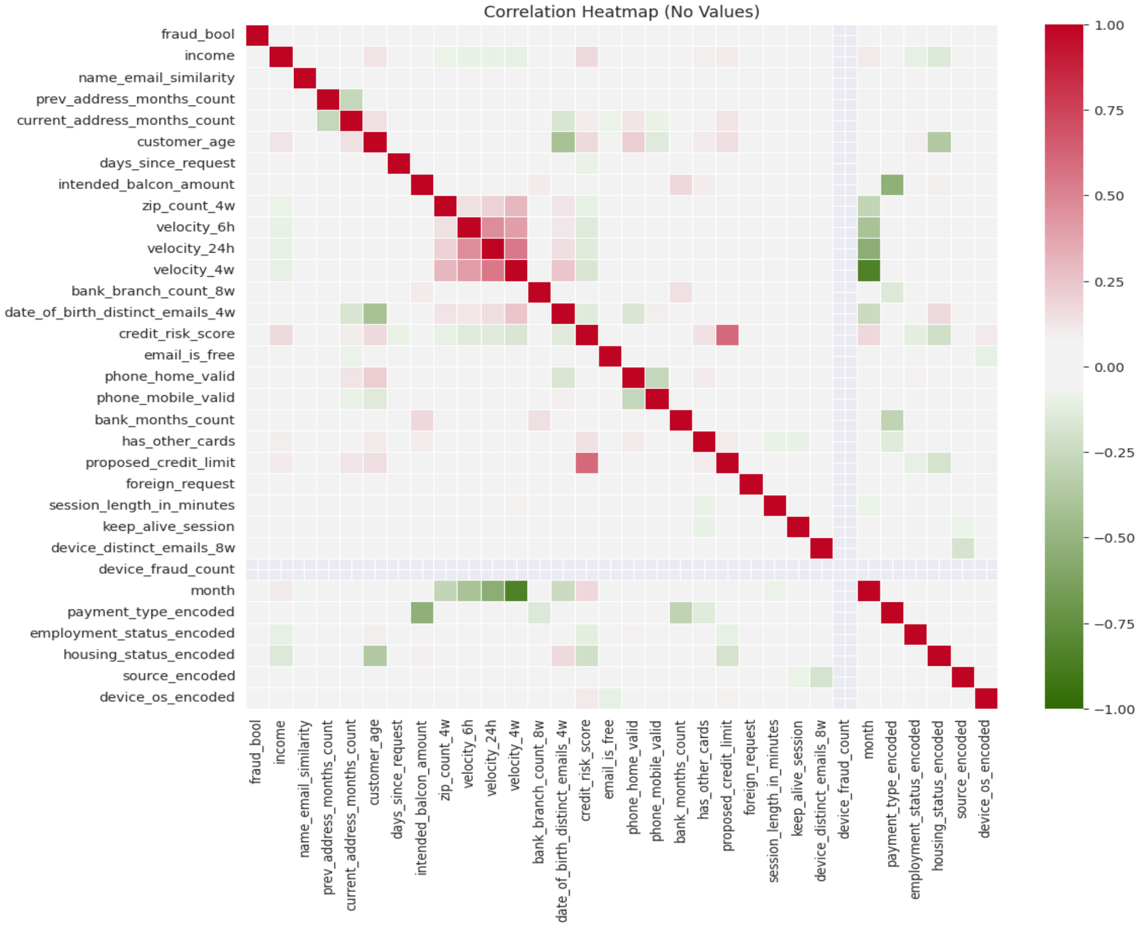


Figure 6: Correlation heatmap of bank account fraud detection.

In Fig. 7, the feature importance values in Bank Account Fraud Detection are very close, especially the top 9 features (name_email_similarity, days_since_request, velocity_4w, credit_risk_score, velocity_6h, session_length_in_minutes, intended_balcon_amount, zip_count_4w, velocity_24h). According to the heat map of the data, the four variables velocity_24h, velocity_4w, velocity_6h, and zip_count_4w have a strong correlation, and they all represent the level of velocity of total applications made. So, the 9 variables represent, the similarity between the applicant's email address and their name, the length of the user's session on the bank's website, the number of days that have passed since the application was made, the application risk internal score, the initial transfer amount, velocity of total applications made these 6 features are the most important.

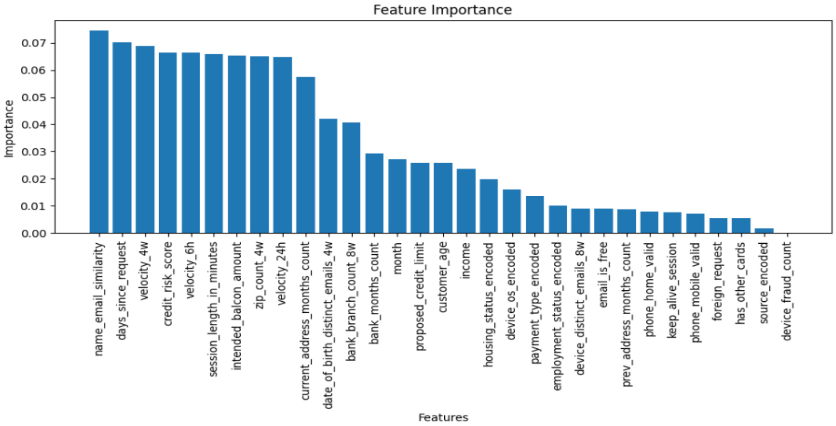


Figure 7: Feature importance of bank account fraud detection.

According to the data analysis, in this dataset, the frequency of fraudulent transactions is high when the similarity between the applicant's email address and their name is below 0.3 (See Fig. 8); the number of days have passed since the application was made in the interval 0-5 (See Fig. 9); the application risk internal score is above 150 (See Fig. 10); or the initial transfer amount is around 100 (See Fig. 11).

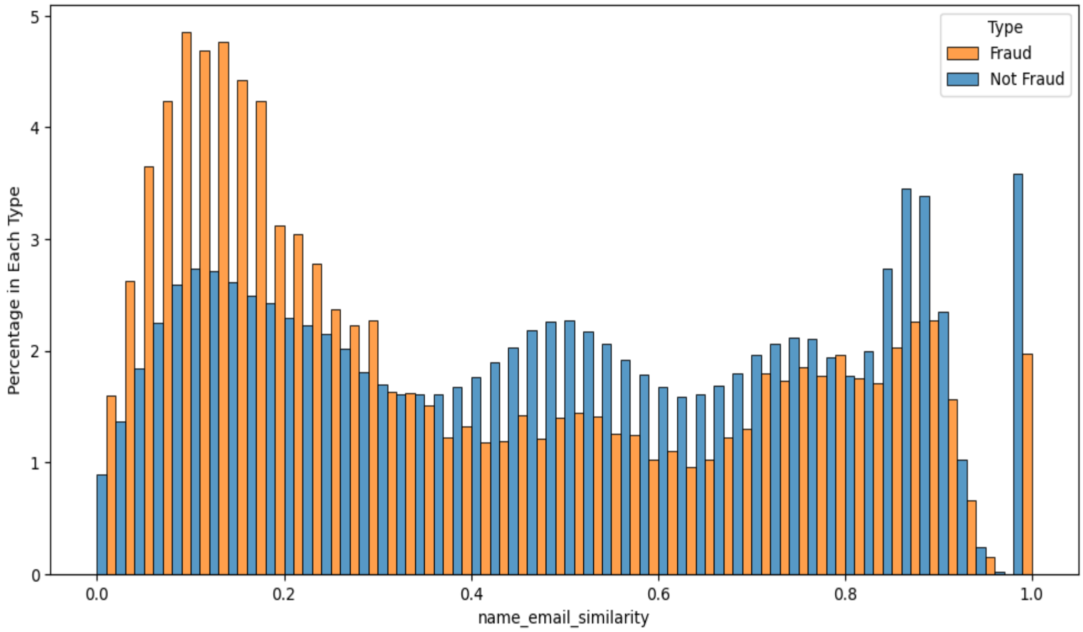


Figure 8: Visualization of similarity between the applicant's email address and their name.

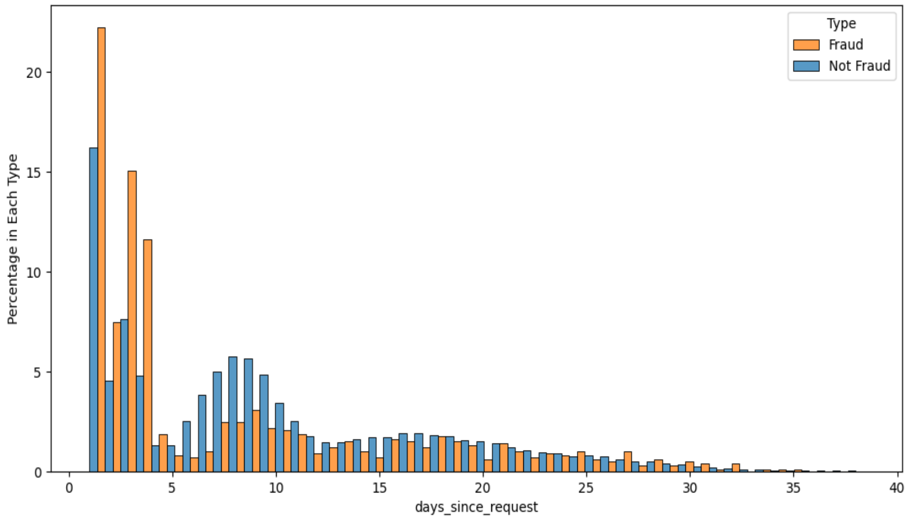


Figure 9: Visualization of number of days have passed since the application was made .

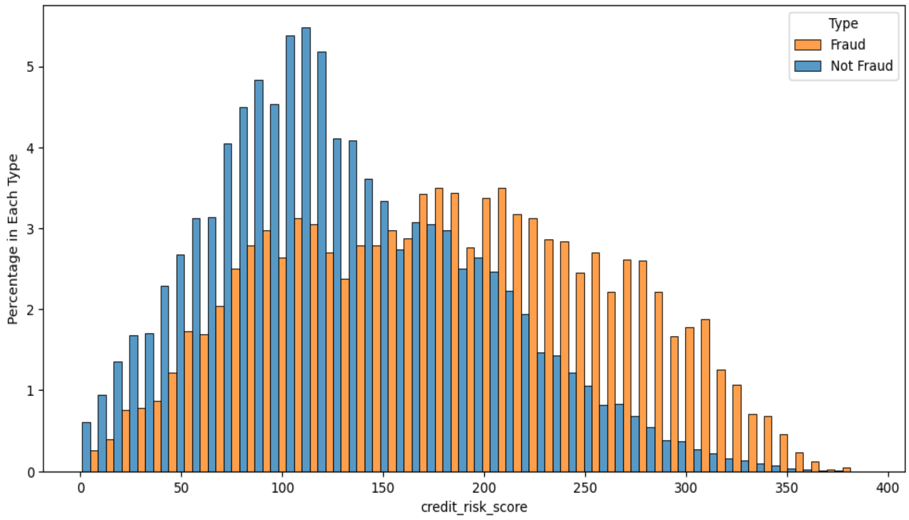


Figure 10: Visualization of application risk internal score.

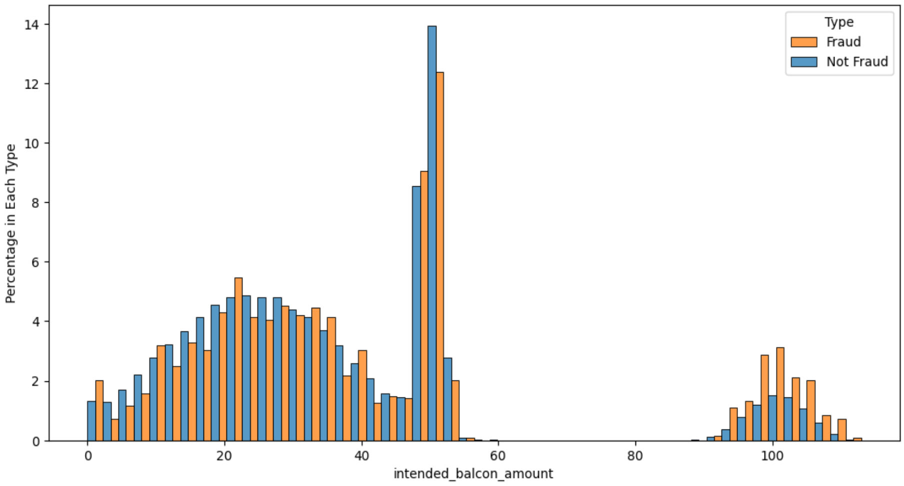
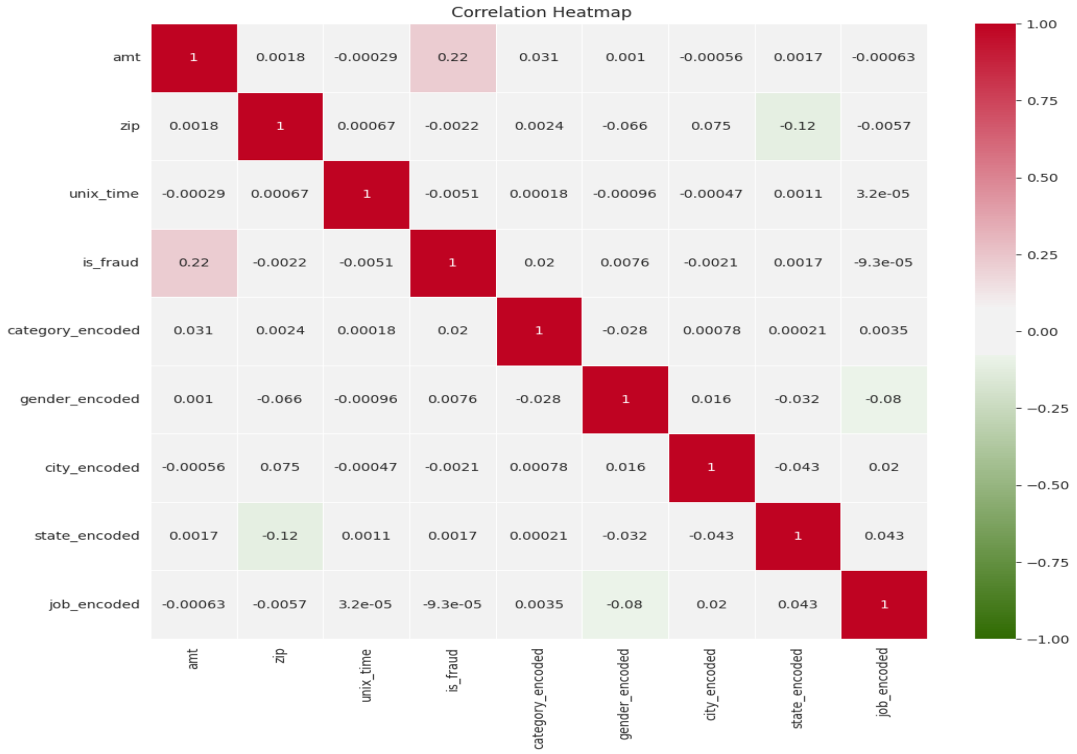


Figure 11: Visualization of the initial transfer amount.

4.3. Credit Card Transactions Fraud Detection

As can be seen from these two plots (See Fig. 12), there is little connection between the variables.



	amt	zip	unix_time	is_fraud	category_encoded	gender_encoded	city_encoded	state_encoded	job_encoded
amt	1	0.0018	-0.00029	0.22	0.031	0.001	-0.00056	0.0017	-0.00063
zip	0.0018	1	0.00067	-0.0022	0.0024	-0.066	0.075	-0.12	-0.0057
unix_time	-0.00029	0.00067	1	-0.0051	0.00018	-0.00096	-0.00047	0.0011	3.2e-05
is_fraud	0.22	-0.0022	-0.0051	1	0.02	0.0076	-0.0021	0.0017	-9.3e-05
category_encoded	0.031	0.0024	0.00018	0.02	1	-0.028	0.00078	0.00021	0.0035
gender_encoded	0.001	-0.066	-0.00096	0.0076	-0.028	1	0.016	-0.032	-0.08
city_encoded	-0.00056	0.075	-0.00047	-0.0021	0.00078	0.016	1	-0.043	0.02
state_encoded	0.0017	-0.12	0.0011	0.0017	0.00021	-0.032	-0.043	1	0.043
job_encoded	-0.00063	-0.0057	3.2e-05	-9.3e-05	0.0035	-0.08	0.02	0.043	1

Figure 12: Correlation heatmap of credit card transactions fraud detection.

Transaction amount is obviously the most important feature in credit card fraud (See Fig. 13). When its value is greater than 200, especially the two intervals of 200-400 and 600-1200, fraudulent transactions are likely to occur (See Fig. 14).

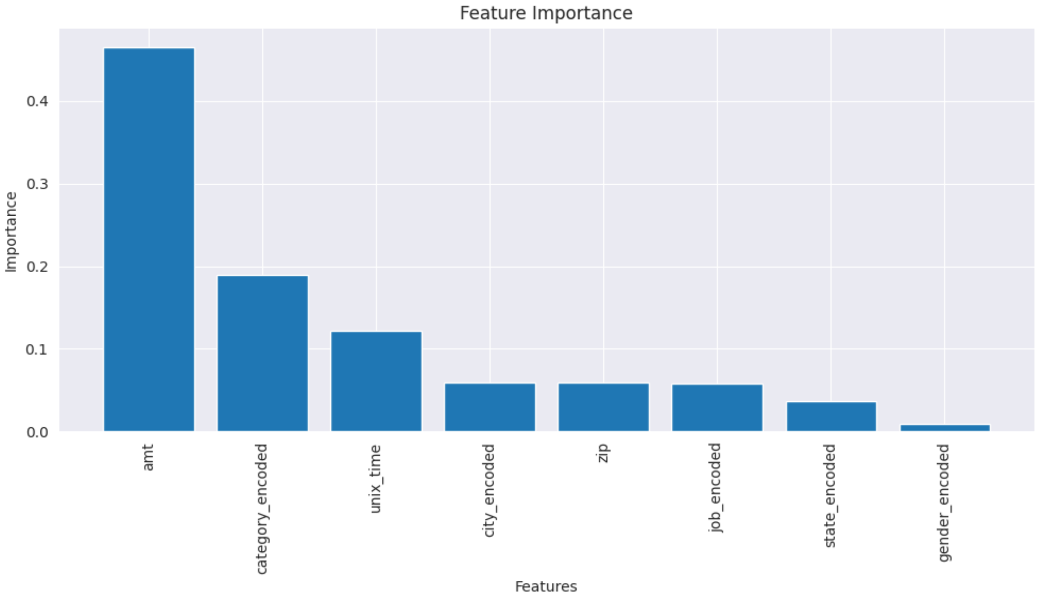


Figure 13: Feature importance of credit card transactions fraud detection.

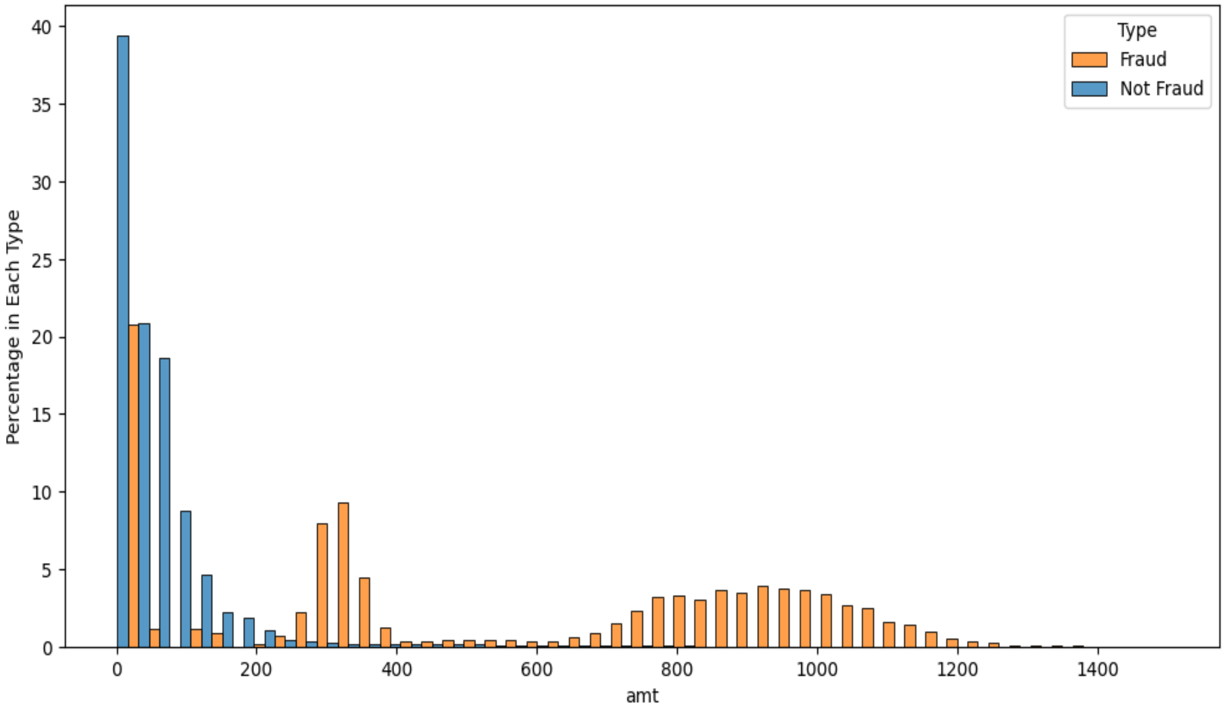


Figure 14: Visualization of transaction amount.

In the time-related analysis, this paper refers to the method of locpham201 in code Fraud Detection using RandomForest+SMOTE+Tuning on Kaggle, net shopping accounted for the highest proportion of fraudulent transactions in the entire timeline (See Fig. 15). In a day, the most common time for fraudulent transactions is from 22 pm to 23 pm, followed by 0-3am (See Fig. 16). This shows that nighttime is the peak period for fraudulent transactions, and net shopping is the most likely transaction type for fraudulent transactions.

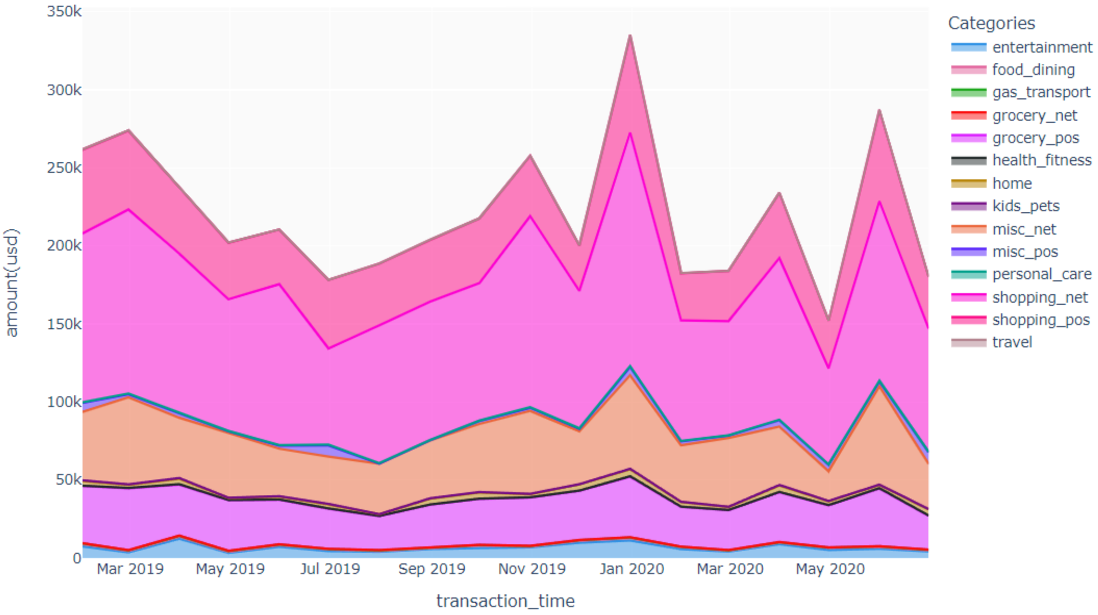


Figure 15: Frequency of different types of fraudulent transactions over time.

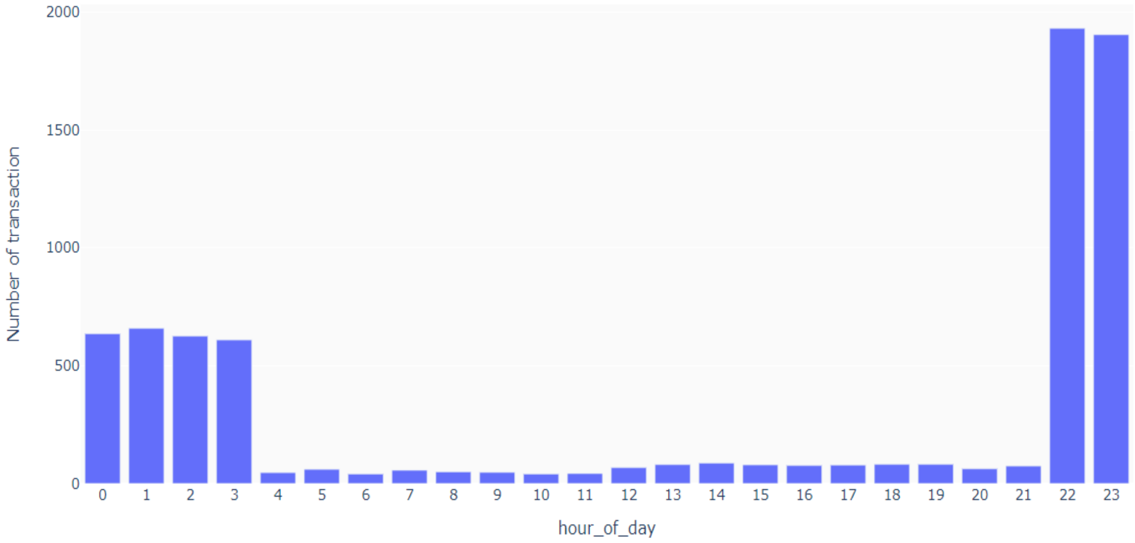


Figure 16: Frequency of fraudulent transactions by time of day.

5. Machine Learning Model: Light GBM

Combining the Fraud Detection knowledge, some features of Light GBM make it ideal for these three scenarios:

Complex Fraud Patterns: Fraud patterns in various scenarios, including online payment fraud, bank account fraud, and credit card transactions fraud, can be complex and dynamic. Light GBM's ability to capture complex interactions and nonlinear relationships in the data makes it suitable for detecting such intricate fraud patterns [8-10]. It can effectively model and identify fraudulent behaviors that may involve multiple features or combinations of features.

High-Dimensional Data: Fraud detection scenarios often involve analyzing a wide range of features related to transactions, customer behavior, account history, and contextual information. Light GBM has excellent performance in processing large-scale structured data sets and has super high training speed, making it suitable for fraud detection in scenarios where numerous features are available. It can efficiently handle a large number of features, which is important for comprehensive fraud analysis.

Class Imbalance: Fraudulent instances are typically rare compared to legitimate transactions, resulting in class imbalance in fraud detection datasets. Light GBM has strong usability, it can handle missing values, outliers, and high cardinality categorical values on features without any special handling. This enables the model to effectively learn from imbalanced data and provide accurate predictions while minimizing false positives.

Efficiency and Scalability: Light GBM uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) technologies. These techniques are designed to significantly improve the efficiency and scalability, making it suitable for large-scale fraud detection scenarios. This efficiency allows for timely fraud detection and real-time decision-making, crucial in scenarios like online payment fraud, bank account fraud, and credit card transactions fraud.

6. Comparison of Conclusions from Different Datasets

On the basis of general fraud detection knowledge, this is the content and key data considered in the detection process in the three scenarios (See Table 4). Combined with this knowledge of fraud detection, important features of different datasets can be analyzed.

Table 4: Content and key data considered in the detection process.

Type of Scenario/Dataset	Key Points of Detection	Key Data Considered
Online Payment Fraud	Protect users, merchants and payment service providers from financial loss and create a safe and secure digital payment environment.	Abnormal transaction amount, risk of payment channel.
Bank Account Opening Fraud	Identify and prevent fraudulent activity by individuals or organized groups attempting to open bank accounts using false identities or stolen personal information for illegal purposes.	Information on applicant's identity, employment, source of funds and intended account use, applicant's address, employment and income, previous banking history, risk assessment.
Credit Card Transactions Fraud	Identifying and preventing unauthorized credit card transactions, thereby protecting cardholders and financial institutions from potential losses.	Abnormal transaction amount, abnormal transaction location, characteristic merchant category, transaction quantity and frequency.

6.1. Online Payment Fraud Detection

Top Important Features: balance before the transaction, the new balance of recipient after the transaction, and the amount of the transaction.

Conclusions from the dataset: High-value Transactions: When the value of the transaction amount is above a certain threshold (e.g., 30,000), fraudulent transactions are more likely to occur. The difference between the old balance of the origin account and the new balance of the recipient after the transaction is indicative of potentially fraudulent activity.

The key data considered in the general knowledge, such as abnormal transaction amount and the risk of payment channels, provides a foundation for identifying potential fraud. These factors highlight the importance of monitoring transactions with unusually high values and assessing the risk associated with different payment channels, which can be useful for initial fraud screening.

Incorporating the conclusions obtained from machine learning in the dataset adds a more specific and data-driven perspective to the fraud detection system. The identification of balance before the transaction, the new balance of the recipient after the transaction, and the amount of the transaction as the top three important features suggests that these variables have a significant impact on predicting fraud.

The reasoning behind these features further enhances the interpretability of the machine learning conclusions. For example, identifying high-value transactions as a significant factor aligns with the general knowledge that abnormal transaction amounts are a key consideration in fraud detection. It confirms that transactions above a certain threshold, such as 30,000, are more likely to be fraudulent.

Additionally, the consideration of balance changes between the origin account and the recipient after the transaction provides valuable insights into potentially fraudulent activity. Significant differences in the account balances indicate suspicious behavior, as fraudsters may attempt to transfer funds to another account or manipulate balances to conceal their activities.

By combining the general knowledge with the specific findings from machine learning, a more comprehensive fraud detection system can be developed. The general knowledge sets the foundation and goals, while the machine learning conclusions provide specific insights into important features and their interpretability. This combination allows for the development of models and algorithms that leverage both human knowledge and data-driven insights, leading to a more accurate and effective fraud detection system.

6.2. Bank Account Fraud Detection

Top Important Features: the similarity between the applicant's email address and their name, the length of the user's session on the bank's website, the number of days that have passed since the application was made, the application risk internal score, the initial transfer amount, velocity of total applications made.

Conclusions from the dataset: the frequency of fraudulent transactions is high when the similarity between the applicant's email address and their name is below 0.3; the number of days have passed since the application was made in the interval 0-5; the application risk internal score is above 150; or the initial transfer amount is around 100.

The key to Bank Account Fraud Detection is the verification of identity and risk. The security system can enhance identity and risk verification and implement a strong identity verification process during account opening or transaction. Additionally, monitoring the amount of money involved in the initial transfer and the rate at which applications are being made can provide insights into potentially fraudulent activity.

By leveraging the identified features, financial institutions can implement the following strategies to enhance their fraud detection capabilities:

Robust Identity Verification: Strengthening identity verification processes by incorporating advanced technologies and techniques can ensure the accuracy and consistency of applicant information, reducing the risk of fraudulent activities.

Real-time Monitoring: Implementing real-time monitoring systems that continuously analyze and assess account opening activities can help detect suspicious patterns and flag potential fraudulent behavior at an early stage, allowing for timely intervention.

Advanced Risk Assessment Algorithms: Developing sophisticated risk assessment algorithms that consider multiple factors, including the identified features from machine learning, can provide a more accurate evaluation of the likelihood of fraudulent account openings. Regular updates and refinement of these algorithms based on emerging fraud patterns can further enhance the effectiveness of the system.

Transaction Monitoring: Establishing thresholds and monitoring the initial transfer amount and the velocity of total applications can enable financial institutions to identify abnormal transaction patterns and detect potential fraud more efficiently. Automated systems can help identify Suspicious activities and trigger additional verification processes.

By incorporating these strategies into their fraud detection systems, financial institutions can improve the accuracy and effectiveness of detecting and preventing fraudulent account openings. This comprehensive approach, combining machine learning insights with human knowledge, enables a more proactive and robust defense against bank account opening fraud.

6.3. Credit Card Transactions Fraud Detection

Top Important Feature: Transaction Amount.

Conclusions from the dataset: Unusual Spending: Unusually high transaction amounts (e.g., above 200) are strong indicators of potentially fraudulent activity. Net Shopping and Timing: Fraudulent transactions are most commonly associated with net shopping scenarios. Nighttime, particularly from 22:00 to 23:00 and 0:00 to 3:00, is the peak period for fraudulent activities, especially in online shopping transactions. The differing feature importance across the three fraud detection scenarios can be attributed to the distinct characteristics of each fraud type and the specific patterns associated with fraudulent behavior. It highlights the importance of tailoring the feature selection and model training process to the unique aspects of each fraud detection domain to achieve accurate and effective fraud detection.

The interpretability of machine learning conclusions can be analyzed in terms of observed relationships in a dataset. The transaction amount plays an important role in detecting fraudulent activity in credit card transactions. Transactions exceeding a transaction amount threshold should be scrutinized to detect and prevent fraud. Fraudulent activity is characterized by both the type of transaction and the timing of the transaction. Security systems can proceed to select, build and train models for these specific transaction types and times.

Based on these findings, the theoretical basis for the development of a more accurate and effective fraud detection system includes:

Prioritizing Transaction Amount: Given that the transaction amount is identified as the most important feature, it should receive significant attention in the development of fraud detection models. Setting appropriate thresholds and closely monitoring high-value transactions can enhance the system's ability to detect and prevent fraud.

Net Shopping Focus: Recognizing the prevalence of fraud in net shopping scenarios, the fraud detection system should be designed to specifically address the risks associated with online transactions. Additional security measures and fraud detection algorithms tailored to online shopping platforms can improve detection accuracy.

Time-based Monitoring: Considering the peak periods of fraudulent activity identified in the dataset, incorporating time-based monitoring mechanisms can enhance the fraud detection system's effectiveness. Paying particular attention to transactions occurring during the identified nighttime periods can help identify suspicious activities.

Continuous Model Improvement: The dataset's emphasis on tailoring the feature selection and model training process highlights the importance of continuous improvement and adaptation. Regularly updating and refining the fraud detection models based on new data and emerging fraud patterns will enable the system to stay ahead of evolving fraudulent techniques.

Combining machine learning with Credit Card Transactions Fraud Detection human knowledge provides a theoretical basis for the development of a more accurate and effective fraud detection system. By incorporating the key findings and focusing on transaction amount, net shopping scenarios, and timing, the system can improve its ability to detect and prevent unauthorized credit card transactions, ultimately protecting cardholders and financial institutions from potential losses.

7. Conclusion

This study explored the applicability of different machine learning models in fraud detection across scenarios such as online payment systems, credit card transactions, and bank accounts and found some improved strategies for machine learning model selection, important features and their interpretation relative to expertise, and improved ability to detect fraud. Light GBM is very suitable for these scenarios because of its ability to capture complex interactions and nonlinear relationships in data, and its ability to deal with unbalanced data problems and high efficiency is very suitable for complex fraud patterns and unbalanced high-dimensional data of fraud detection problems. By combining machine learning techniques with human knowledge through feature engineering and domain expertise, the interpretability of fraud detection systems was enhanced. Important features were identified for each domain, such as transaction amounts, balance changes, and application ant characteristics, providing insights into potentially fraudulent activities. The combination of general fraud detection knowledge and machine learning conclusions allowed for the development of more accurate and effective fraud detection systems. Strategies such as robust identity verification, real-time monitoring, advanced risk assessment algorithms, and transaction monitoring can be implemented based on the identified features to enhance fraud detection capabilities.

References

[1] Aladakatti, D., Gagana, P., Kodipalli, A., Kamal, S.: Fraud detection in Online Payment Transaction using Machine Learning Algorithms. In 2022 International Conference on Smart and Sustainable Technologies in Energy and Power Sectors, 223-228 (2022).

[2] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., Ahmed, M.: Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access 10, 39700-39715 (2022).

[3] Kolodiziev, O., Mints, A., Sidelov, P., Pleskun, I., Lozynska, O.: Automatic machine learning algorithms for fraud detection in digital payment systems. Восточно-Европейский журнал передовых технологий 5(9-107), 14-26 (2020).

[4] Itri, B., Mohamed, Y., Mohammed, Q., Omar, B.: Performance comparative study of machine learning algorithms for automobile insurance fraud detection. Third International Conference on Intelligent Computing in Data Sciences, 1-4 (2019).

[5] Awoyemi, J. O., Adetunmbi, A. O., Oluwadare, S. A.: Credit card fraud detection using machine learning techniques: A comparative analysis. International conference on computing networking and informatics, 1-9 (2017).

[6] Lin, K., Gao, Y.: Model interpretability of financial fraud detection by group SHAP. Expert Systems with Applications 210, 118354 (2022).

[7] Kaggle, https://www.kaggle.com/, last accessed 2023/7/1.

[8] Al Daoud, E.: Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering 13(1), 6-10 (2019).

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

[10] Li, K., Xu, H., Liu, X.: Analysis and visualization of accidents severity based on LightGBM-TPE. Chaos, Solitons & Fractals 157, 111987 (2022).