Course Project Report

Credit Card Fraud Detection using ML and DS

Submitted By

Shashank Reddy Muppidi (211AI033)
Chaithanya Swaroop(211AI010)

as part of the requirements of the course

Data Science (IT258) [Feb - Jun 2023]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

FEB-JUN 2023

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Course project Work Report entitled **“Credit Card Fraud Detection using ML and DS”** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Shashank Reddy Muppidi	211AI033	M.Shashank 12/06/23
Chaithanya Swaroop	211AI010	Chaithanya Swaroop 12|06|2023

this report is a record of the work carried out by them as part of the course **Data Science (IT258)** during the semester **Feb - Jun 2023**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

DECLARATION

We hereby declare that the project report entitled **"Credit Card Fraud Detection using ML and DS"** submitted by us for the course **Data Science (IT258)** during the semester **Feb-Jun 2023**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Shashank Reddy Muppidi	211AI033	M.Shashank 12/06/23
2. Chaithanya Swaroop	211AI010	Chaithanya Swaroop 12|06|2023

Place: NITK, Surathkal
Date: 12th June 2023

Credit Card Fraud Detection using ML and DS

Shashank Reddy Muppidi - Team 2, 211AI033 Chaithanya Swaroop - Team 2, 211AI010

***Abstract*— Given the growing occurrence of credit card fraud within the financial industry, there is a pressing demand for effective systems to detect fraudulent activities. This project introduces a comprehensive approach to credit card fraud detection by leveraging machine learning techniques. Our objective is to create a system that can accurately identify fraudulent transactions while minimizing false positives, thereby enhancing the overall security and reliability of credit card transactions. Initially, we undertake thorough feature engineering to extract pertinent information from transaction data, including transaction amount, time, and other derived features. Subsequently, we employ various machine learning algorithms, such as Logistic Regression, Local Outlier Factor, and Isolation Forest, to train and evaluate our models using a labeled dataset consisting of both genuine and fraudulent credit card transactions. This dataset enables us to effectively capture patterns and anomalies associated with fraudulent activities. It is imperative for credit card banks to identify fraudulent transactions to prevent customers from being charged for unauthorized purchases. Data science, in conjunction with machine learning, proves highly effective in addressing such issues. Our project aims to demonstrate the application of machine learning in modeling a dataset for credit card fraud detection. The problem involves developing a model based on past instances of fraudulent credit card transactions, which is then utilized to detect potentially fraudulent transactions in real-time. Our primary objective is to identify fraudulent transactions while minimizing false classifications. The process involves analyzing and preprocessing the dataset, applying multiple anomaly detection algorithms (such as the local outlier factor and the isolation forest) to the credit card transaction data transformed using principal component analysis (PCA).**

***Keywords:* Detection Techniques, Credit Card Fraud Detection, Fraudulent Transactions, k Nearest Neighbors, Local Outlier Factor**

I. Introduction

Unfortunately, the widespread usage of credit cards as a convenient payment method has led to increased credit card fraud cases. Fraudulent transactions result in substantial financial losses for financial institutions and consumers, eroding trust in electronic transaction security. Robust fraud detection systems are essential to mitigate these risks and ensure the integrity of credit card transactions.

Preventing credit card fraud is imperative to safeguard the account owner and their card issuer against unauthorized usage. Necessary precautions must be taken to minimize and eliminate fraudulent behavior. Ensuring that the account owner and issuer are always aware of any unauthorized activity is crucial to avoid potential losses. As part of fraudulent detection, we monitor user activity to identify and prevent any fraudulent behavior, such as intrusions, late payments, and other aggressive actions. Our goal is to maintain a safe and fair environment for all users. Effective fraudulent detection requires vigilant monitoring of user activities to swiftly recognize and prevent suspicious behavior, ranging from intrusive actions to delinquent payments and other forms of aggressive conduct. Detecting fraud is an intricate and challenging undertaking, mainly when dealing with factors such as class imbalance that can complicate the process. With the vast majority of valid transactions, fraudulent ones can be difficult to spot, significantly since transaction patterns can change frequently. Nevertheless, implementing a fraud detection system is achievable. Automated tools scan a high volume of payment requests and use machine learning algorithms to analyze and flag suspicious transactions for review by experts. These transactions are categorized as either Online or Offline frauds, and experts will contact cardholders to confirm the transaction's authenticity. Over time, this process helps improve the system's ability to detect fraud, and we are confident in the system's capacity to do so effectively.

Figure-1 depicts the flowchart of the transactions of the two main algorithms.

II. LITERATURE

This refers to engaging in intentional actions that violate laws, regulations, or policies for personal financial gain. A considerable amount of literature on fraud detection in this domain has been published.
Kho et al. [1] explore the detection of credit card

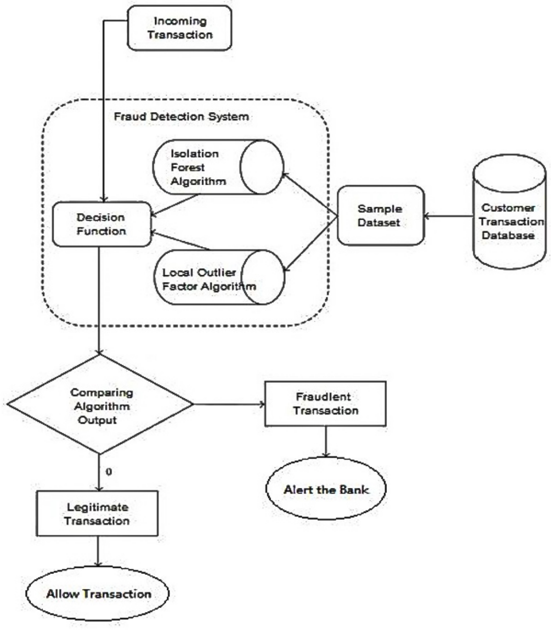


Fig. 1. Flowchart

fraud by analyzing transaction behavior. Their study addresses the growing concern of credit card fraud and proposes a methodology that focuses on transaction behavior to identify fraudulent activities. The authors provide an overview of the prevalence of credit card fraud and its negative impact on individuals and financial institutions. They then conduct a comprehensive literature review of existing techniques and approaches used in credit card fraud detection. The review covers various methods, including rule-based systems, anomaly detection, and machine learning algorithms. The limitations of rule-based systems are highlighted, as they rely on predefined rules and thresholds that may not adapt well to evolving fraud patterns. Anomaly detection approaches are also discussed, which can detect unusual patterns but struggle with distinguishing between genuine anomalies and fraudulent activities. Machine learning algorithms, particularly supervised learning methods, are identified as a promising approach for fraud detection. Different algorithms such as decision trees, support vector machines (SVM), and neural networks are discussed, along with their strengths and weaknesses in credit card fraud detection. The paper emphasizes the importance of feature selection and engineering in building effective fraud detection models, suggesting transaction attributes like amount, location, time, and merchant category as potential features to capture behavioral patterns indicative of fraud.

Clifton et al. [2] and Suman et al. [3] have conducted extensive research in this field, exploring techniques such as data mining applications, automated fraud detection, and adversary detection. While these methods and algorithms have shown unexpected success in some areas, they have not provided a consistent and robust solution to fraud detection. Similar research has been conducted by Wen-Fang et al. [4], who utilized outlier detection mining and distance summing algorithms in credit card emulation experiments to detect fraudulent transactions. They used attributes of customer behavior to calculate the distance between observed values and expected values, and unconventional techniques like z-groups were found to be effective for medium-sized online transactions. Efforts are also being made to improve the interaction of alert feedback in the event of fraudulent transactions.

Zanin et al. [5] propose a novel approach for credit card fraud detection using paren clitic network analysis. They highlight the significant financial losses caused by credit card fraud and the necessity for effective fraud detection systems. Dal Pozzolo et al. [6] present a realistic modeling approach and a novel learning strategy for credit card fraud detection. They address the challenge by proposing an innovative approach that combines realistic modeling and a learning strategy to handle imbalanced class distributions often observed in real-world datasets.

Overall, these studies contribute to the existing body of knowledge in credit card fraud detection, offering insights into transaction behavior analysis, machine learning algorithms, and novel approaches to improve detection accuracy and address class imbalance challenges.

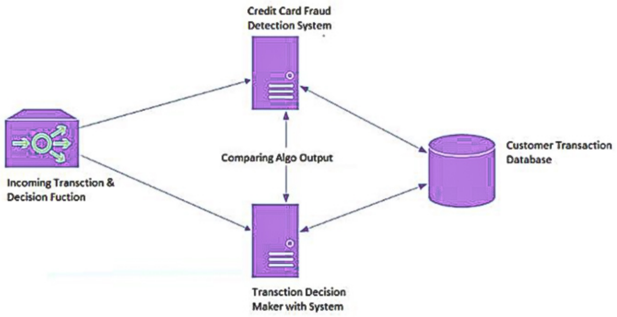

Credit Card Fraud Detection System
Incoming Transction & Decision Fuction
Comparing Algo Output
Customer Transaction Database
Transction Decision Maker with System

Fig. 2. Rough Architecture Diagram

III. METHODOLOGY

This study proposes utilizing cutting-edge machine learning techniques to detect unusual behaviors, often referred to as outliers. An example of the basic architecture diagram can be seen in Figure-2. For a more comprehensive understanding, Figure-3 presents the complete architecture diagram, incorporating real-life components. The columns in the diagram represent Time, Amount, and Class. Time indicates the duration between consecutive transactions, while Amount refers to the monetary value transferred. Class 0 corresponds to legitimate transactions, whereas Class 1 denotes fraudulent transactions. To visually examine the dataset and identify any inconsistencies, various graphs have been plotted. Figure-4 illustrates the significant disparity between the number of fraudulent transactions and legitimate ones. Additionally, Figure-5 showcases the distribution of the total transaction amounts.

The majority of transactions are of small magnitude, with only a limited number approaching the maximum allowable exchange amount. After validating the dataset, we generate a histogram for each individual entry. The purpose of this is to create a visual representation of the dataset, enabling us to identify any missing values. This step ensures that the machine learning algorithms can handle the dataset effectively, without requiring imputation for missing values. Subsequently, Figure-6 depicts a heatmap to visualize the data using color, allowing us to analyze the relationship between the class variable and the predictors that yield the output. In Figure 7, the percentage counts of fraudulent and non-fraudulent transactions are presented. This figure provides a visual representation of the distribution of fraudulent and non-fraudulent transactions in the dataset, allowing us to observe the relative proportions of each category. Figure 8 displays a boxplot for the comparison of transaction amounts. The boxplot provides a summary of the distribution of transaction amounts for both fraudulent and non-fraudulent transactions. It allows us to compare the median, quartiles, and potential outliers between the two categories, giving insights into any significant differences in transaction amounts. Figure 9 showcases the covariance analysis results. This figure presents the covariance matrix or covariance heatmap, which illustrates the relationships and dependencies between different attributes or variables in the dataset. By examining the covariance values, we can gain insights into how the attributes co-vary and potentially identify patterns or correlations. Figure 10 demonstrates the application of principal component analysis (PCA). This figure provides a visualization of the dataset after applying PCA, showing the reduced dimensions and the variance explained by each principal component.

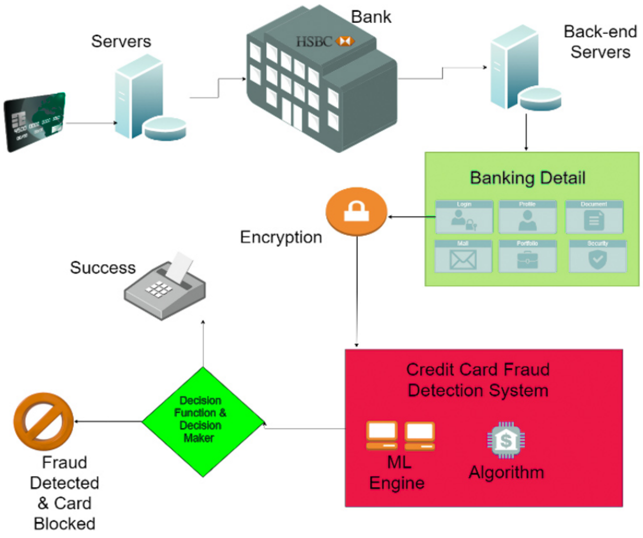

Servers
Bank
HSBC
Back-end Servers
Banking Detail
Encryption
Success
Decision Function & Decision Maker
Fraud Detected & Card Blocked
Credit Card Fraud Detection System
ML Engine
Algorithm

Fig. 3. Complete Architecture Diagram

Now that the dataset has undergone processing and formatting, certain modifications have been made to ensure fair grading. The Class column has been removed, and the time and amount columns have been standardized. A set of algorithms from various modules is used to process the data. Once the data has been fitted into a model,

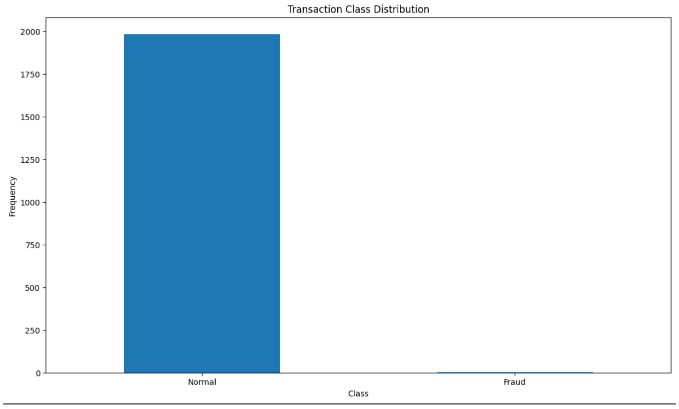


Fig. 4. Value counts of Fraudulent and Non-Fraudulent transactions

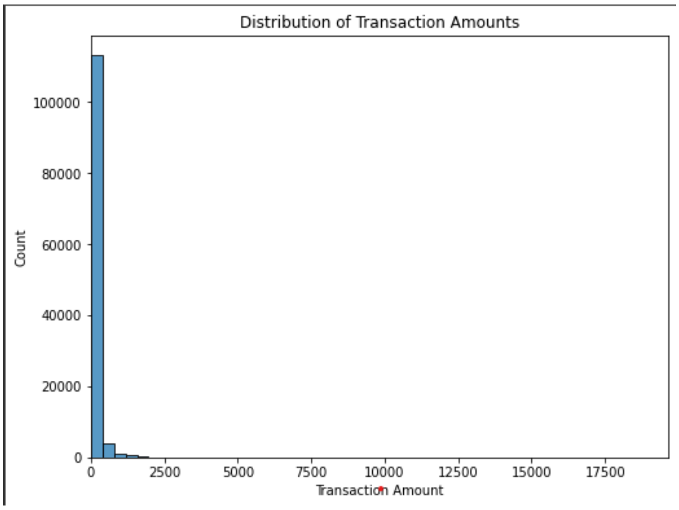


Fig. 5. Total amount distribution

outlier identification modules like Isolation Forest Algorithm, Local Outlier Factor, and Logistic Regression are applied to identify any anomalies.

IV. IMPLEMENTATION

Implementing this concept in practicality presents challenges since it necessitates collaboration from financial institutions. However, banks are reluctant to disclose information due to their competitive nature in the market, as well as legal obligations and the need to safeguard their users' data.

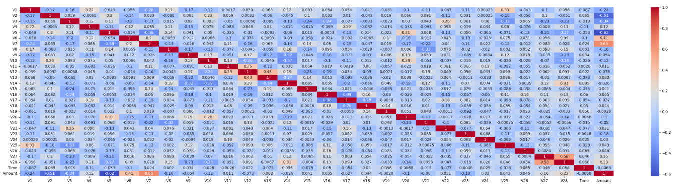


Fig. 6. Correlation Analysis

Exploratory Data Analysis (EDA) using visualization techniques was conducted to gain insights into the data and identify patterns or anomalies. The primary objective of EDA was to understand the distribution and characteristics of the dataset, as well as to detect any potential relationships or trends that could assist in the identification of fraudulent transactions. To begin the EDA process, various visualization methods were employed. Histograms were created to examine the distribution of numerical features such as transaction amount. This allowed us to assess the presence of any outliers or skewed distributions that might indicate fraudulent activity. Box plots were also generated to visualize the quartiles, median, and potential outliers within the dataset. Additionally, scatter plots were utilized to investigate the relationships between different variables. This enabled us to explore any potential correlations or patterns that might exist between these variables and fraudulent transactions. Furthermore, bar charts and pie charts were employed to visualize categorical features, such as the transaction type or merchant category. This allowed us to identify any dominant categories or transaction types associated with fraudulent activities.

Data Preprocessing steps were applied, followed by the implementation of various machine learning algorithms. The objective of the data preprocessing stage was to clean, transform, and normalize the data, ensuring its suitability for modeling purposes. First, missing values were examined and handled appropriately using imputation. Next, feature scaling was applied to the numerical features V1, V2, ..., V28, Time, and Amount. This process ensured that all numerical features were on a comparable scale, preventing any particular feature from dominating the modeling process. Furthermore, as the dataset had an imbalanced distribution between fraudulent and non-fraudulent transactions, sampling techniques were applied to address this issue. Specifically, oversampling the minority class (fraudulent transactions) using the ”SMOTE” was performed. ”SMOTE” generated synthetic instances of the minority class

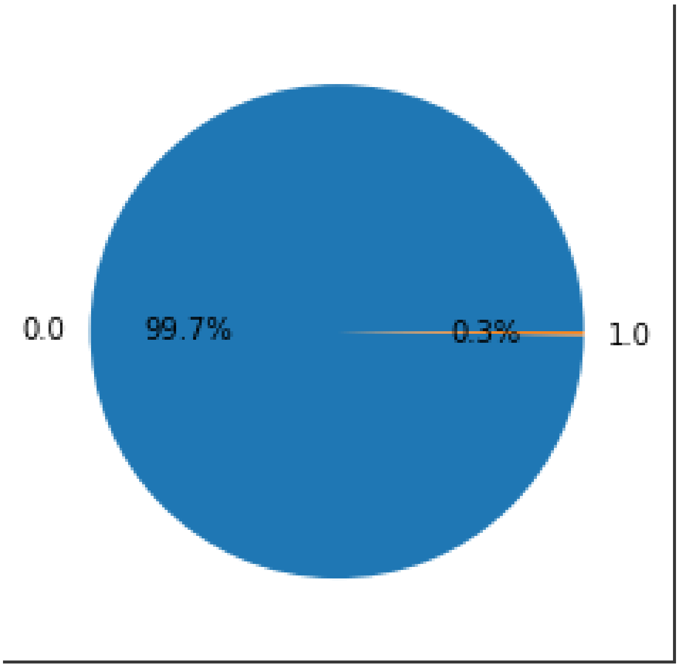


Fig. 7. Percentage counts of Fraudulent and Non-Fraudulent transactions

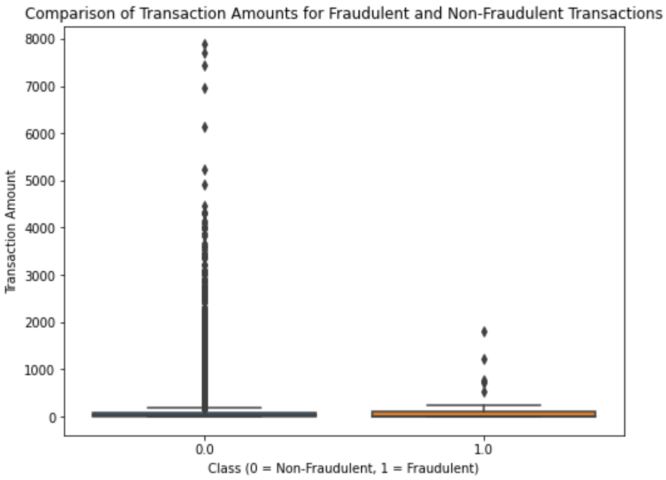


Fig. 8. Boxplot for Comparision of Transaction Amount

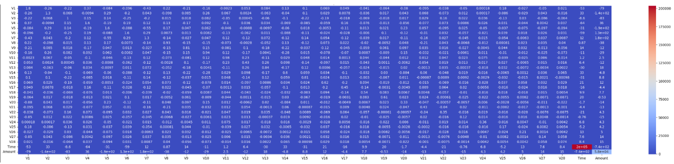


Fig. 9. Covariance Analysis

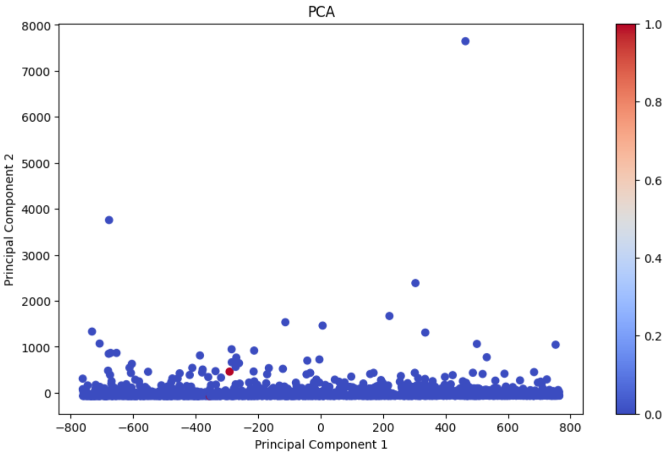
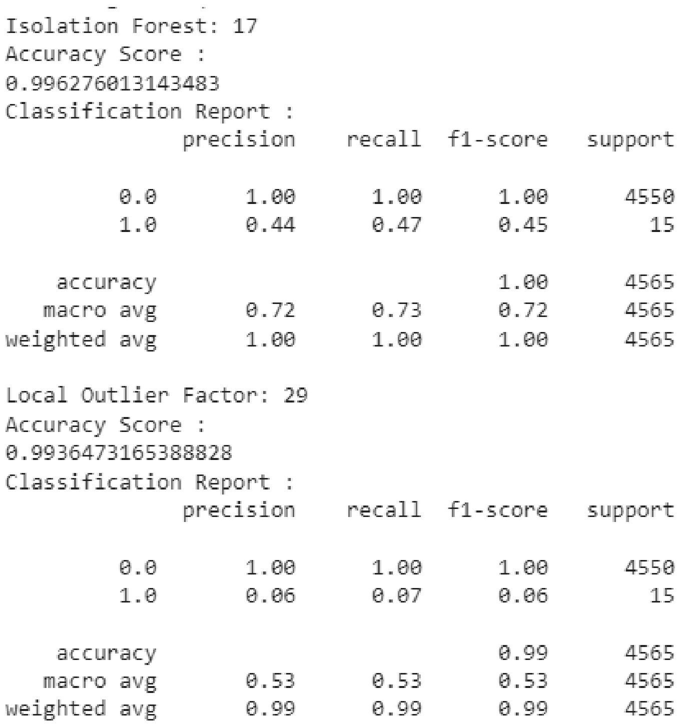


Fig. 10. Applying PCA

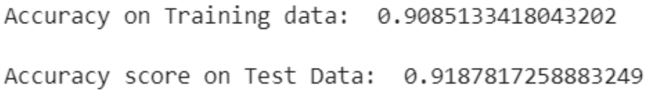
to balance the class distribution, thus mitigating potential bias during model training. After the data preprocessing stage, the dataset we splitted into testing and training subsets, with the conventional 80:20 ratio. The training set was used for model training, while the testing set was employed for evaluating the performance of the trained models.

Advanced analysis techniques were employed to gain deeper insights and enhance the predictive capabilities of the models. This involved conducting correlation analysis, covariance analysis, dimensionality reduction, and feature engineering. Correlation analysis was performed to explore the relationships between the different features in the dataset. Specifically, the Pearson correlation coefficient was calculated for each pair of numerical features (V1, V2, ..., V28, Time, Amount) to measure the strength and direction of the linear relationship between them. This analysis helped identify potential correlations that could be indicative of fraudulent transactions or provide additional predictive power to the models. Covariance analysis was also conducted to examine the statistical relationship between pairs of variables. It enabled the assessment of how changes in one variable may be associated with changes in another variable. By analyzing the covariance matrix of the numerical features, patterns and dependencies within the dataset were identified, aiding in the identification of potential discriminatory factors. To address the challenge of larger dimensionality in the



```
Isolation Forest: 17
Accuracy Score :
0.996276013143483
Classification Report :
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      4550
         1.0       0.44      0.47      0.45        15

    accuracy                           1.00      4565
   macro avg       0.72      0.73      0.72      4565
weighted avg       1.00      1.00      1.00      4565

Local Outlier Factor: 29
Accuracy Score :
0.9936473165388828
Classification Report :
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      4550
         1.0       0.06      0.07      0.06        15

    accuracy                           0.99      4565
   macro avg       0.53      0.53      0.53      4565
weighted avg       0.99      0.99      0.99      4565
```

Fig. 11. Analysis of different models



```
Accuracy on Training data:  0.9085133418043202

Accuracy score on Test Data:  0.9187817258883249
```

Fig. 12. Analysis using Logistic Regression Method

dataset, dimensional reduction were employed. PCA was applied to reduce the dimensionality of the numerical features while retaining as much information as possible. This technique transformed the normal set of features into uncorrelated features. By retaining only the most significant principal components, the dataset's dimensionality was reduced, facilitating more efficient modeling while preserving the relevant information. Additionally, feature engineering was performed to create new features or transform existing ones, with the goal of capturing relevant patterns and improving the predictive power of the models. Feature engineering techniques like scaling or encoding were applied to create informative and meaningful features that better represented the underlying patterns in the data.

Model	Accuracy
Isolation Forest	99.6
Local Outlier Factor	99.3
Logistic Regression	90.8

V. CONCLUSION AND FUTURE ENHANCEMENTS

Undoubtedly, engaging in fraudulent activities with a credit card is a dishonest and criminal act. This article presents the most common fraud schemes, along with techniques for identifying them, while also discussing recent research in this field. we provided an algorithm, pseudocode, implementation details, and experimental results. Although the algorithm achieves an accuracy rate exceeding 99.6 as shown in Figure-11, its precision remains relatively low when considering only a tenth of the dataset. Furthermore, the precision increases only slightly when the entire dataset is used. This high accuracy is seen due to the significant difference between the number of respective transaction types.

If this research were to be applied commercially, a small portion of the data could be made accessible. Over time, the program would become more effective as it receives additional data, thanks to its reliance on machine learning techniques. We have successfully developed a system that can come very close to achieving the desired outcome with sufficient time and data. There is still room for improvement, as is the case with any project of this nature.

To further enhance this model, other algorithms can be incorporated, provided their output format aligns with the existing ones. As demonstrated in the code, adding these modules is straightforward once this requirement is met. The dataset holds potential for further growth. As previously shown, the precision of the algorithms improves with an increase in dataset size. Expanding the dataset by acquiring more data will undoubtedly improve the model's capacity to detect fraud and minimize false positives. However, it is vital to secure legal support from the banks themselves in order to pursue this expansion effectively.

REFERENCES

[1] Kho, John Richard D., and Larry A. Vea. ”Credit card fraud detection based on transaction behavior.” TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017.

[2] Phua, Clifton, et al. ”A comprehensive survey of data mining-based fraud detection research.” arXiv preprint arXiv:1009.6119 (2010).

[3] Suman, Mitali Bansal, and Mitali Bansal. ”Survey paper on credit card fraud detection.” International Journal of Advanced Research in Computer Engineering Technology (IJARCET) 3.3 (2014): 827-832.

[4] Yu, Wen-Fang, and Na Wang. ”Research on credit card fraud detection model based on distance sum.” 2009 International Joint Conference on Artificial Intelligence. IEEE, 2009.

[5] Zanin, Massimiliano, et al. ”Credit card fraud detection through parenclitic network analysis.” Complexity 2018 (2018).

[6] Dal Pozzolo, Andrea, et al. ”Credit card fraud detection: a realistic modeling and a novel learning strategy.” IEEE transactions on neural networks and learning systems 29.8 (2017): 3784-3797.

[7] Rathore, Anjali Singh, et al. ”Credit Card Fraud Detection using Machine Learning.” 2021 10th International Conference on System Modeling Advancement in Research Trends (SMART). IEEE, 2021.

[8] Weston, David J., et al. ”Plastic card fraud detection using peer group analysis.” Advances in Data Analysis and Classification 2 (2008): 45-62.

Document Viewer

Turnitin Originality Report

Processed on: 12-Jun-2023 06:36 IST
ID: 2113973451
Word Count: 2787
Submitted: 1

Team2_ChaithanyaSwaroop_ShashankReddy_project...
By Anonymous

Similarity Index	Similarity by Source	
10%	Internet Sources:	5%
	Publications:	5%
	Student Papers:	8%

include quoted | include bibliography | exclude small matches | mode: quickview (classic) report

print | refresh | download

2% match (student papers from 10-May-2023)
Submitted to University of East London on 2023-05-10 ☒

1% match (Internet from 08-Jan-2022)
https://www.ijert.org/credit-card-fraud-detection-using-machine-learning-and-data-science ☒

1% match (student papers from 23-Jun-2022)
Submitted to Visvesvaraya Technological University, Belagavi on 2022-06-23 ☒

1% match (student papers from 07-Nov-2022)
Submitted to SRM University on 2022-11-07 ☒

1% match (student papers from 10-May-2023)
Submitted to University of North Texas on 2023-05-10 ☒

1% match (student papers from 29-Nov-2022)
Submitted to BPP College of Professional Studies Limited on 2022-11-29 ☒

1% match (student papers from 02-Nov-2022)
Submitted to New Jersey Institute of Technology on 2022-11-02 ☒

<1% match (Internet from 16-Oct-2022)
https://www.ijert.org/research/credit-card-fraud-detection-using-machine-learning-IJERTCONV7IS10036.pdf ☒

<1% match (student papers from 07-Jun-2023)
Submitted to Federation University on 2023-06-07 ☒

<1% match (student papers from 12-Apr-2023)
Submitted to Liverpool John Moores University on 2023-04-12 ☒

<1% match (student papers from 06-Jan-2023)
Submitted to University of Sunderland on 2023-01-06 ☒

<1% match (Swarnalatha K S, Krishna Kumar Shah, Kishore Kumar, Krishna Kumar Patel, Aashutosh Raj Sah. "Credit Card Fraud Detection Using Machine Learning Model", 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022)
Swarnalatha K S, Krishna Kumar Shah, Kishore Kumar, Krishna Kumar Patel, Aashutosh Raj Sah. "Credit Card Fraud Detection Using Machine Learning Model", 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022 ☒

<1% match (Internet from 26-Sep-2021)
https://www.ijeat.org/wp-content/uploads/Souvenir_Volume-9_Issue-3_Februry_2020.pdf ☒

<1% match (Internet from 08-Oct-2022)
https://ijsrst.com/paper/9580.pdf ☒

<1% match (Akanksha Bansal, Hitendra Garg. "An Efficient Techniques for Fraudulent detection in Credit Card Dataset: A Comprehensive study", IOP Conference Series: Materials Science and Engineering, 2021)
Akanksha Bansal, Hitendra Garg. "An Efficient Techniques for Fraudulent detection in Credit Card Dataset: A Comprehensive study", IOP Conference Series: Materials Science and Engineering, 2021 ☒