EXPLORATORY DATA ANALYSIS (EDA)
ON
CREDIT CARD FRAUD DETECTION

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

SHAHANA T
12219504
RK22URA26

Supervisor

VED PRAKASH CHAUBEY

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2024

1. Supervisor’s certificate

This is to certify that the work presented in the **B.Tech** Dissertation/Dissertation Proposal titled “**EDA Project on Credit Card Fraud Dataset**,” submitted by **Shahana T** at **Lovely Professional University, Phagwara, India**, is a genuine record of her original work conducted under my guidance. This work has not been submitted for any other degree elsewhere.

Signature of Supervisor

(Name of Supervisor)

Date:

2. Acknowledgment

I am deeply grateful to **Er. Ved Prakash Chaubey sir** for his invaluable guidance, support, and encouragement throughout the development of this project, *"***EDA Project on Credit Card Fraud Dataset***."* His insights and expertise have been instrumental in shaping my understanding and approach toward this work.

I would also like to thank **Lovely Professional University** for providing the platform and resources necessary to carry out this project. The access to data, tools, and mentorship has been crucial to my learning experience.

A special note of gratitude goes to my family and friends, whose unwavering support and motivation have inspired me to complete this work successfully.

Lastly, I extend my heartfelt thanks to everyone who directly or indirectly contributed to the successful completion of this project. Your contributions have been deeply appreciated.

Shahana T

Table of content

1	***Supervisor's certificate***
2	***Acknowledgement***
3	***Abstract***
4	***Problem Statement and Data Description***
5	***Solution Approach***
6	***Libraries***
7	***Introduction***
8	***Literature Review***
9	***Methodology***
10	***Results***
11	***Analysis***
12	***Conclusion***
13	***References***
14	***GitHub Repository Link***

3. Abstract

This analysis aims to identify and understand fraudulent credit card transactions using a dataset of 284,807 transactions carried out by European cardholders over a two-day period in September 2013. The dataset is highly imbalanced, with fraudulent transactions making up only 0.172% of the total, presenting unique challenges that require specialized handling to ensure unbiased conclusions. The 31 features in the dataset, most derived using Principal Component Analysis (PCA), exhibit no direct multicollinearity, which simplifies interpretation and facilitates the identification of patterns. Through Exploratory Data Analysis (EDA), distinct behaviors of fraudulent transactions were revealed, such as higher variance in specific transaction amounts and anomalies that deviate significantly from typical patterns. These anomalies could represent critical edge cases essential for improving fraud detection systems. Additionally, the analysis highlights the need for tailored techniques to address the class imbalance, ensuring accurate modeling and reducing bias. By leveraging these insights, financial institutions can strengthen their fraud detection systems, mitigate financial losses, and enhance customer protection. This study underscores the pivotal role of effective data visualization and statistical methods in illuminating actionable trends, thereby enabling the development of robust machine learning models that balance fraud detection accuracy with minimized false positives. The findings not only contribute to enhancing financial security but also emphasize the importance of continuous monitoring and refinement of detection systems in the dynamic landscape of fraud.

4. Problem statement and data description

Credit card fraud poses a significant challenge for financial institutions, as it results in financial losses and erodes customer trust. Detecting fraudulent transactions is critical to safeguarding both the banks and their customers. The primary goal of this project is to identify patterns and anomalies in fraudulent credit card transactions using Exploratory Data Analysis (EDA). This analysis aims to provide insights that can inform machine learning models and help prevent fraud effectively.

Dataset Description

The dataset used for this project contains transactional data from European credit cardholders over a two-day period in September 2013. Key details of the dataset are as follows:

- **Number of Records**: 284,807 transactions.
- **Number of Features**:
  - 31, including 28 features generated through Principal Component Analysis (PCA).
  - <u>Time</u>: Seconds elapsed between the transaction and the first transaction in the dataset.

- Amount: Transaction amount in euros.
- Class: Target variable indicating whether a transaction is fraudulent (1) or legitimate (0).

- **Class Distribution**:
  - Fraudulent transactions account for only **0.172%** of the data.
  - Legitimate transactions represent the remaining **99.828%**.

The dataset's high imbalance highlights the challenge of distinguishing fraud cases from legitimate transactions, necessitating advanced analysis and careful handling to derive actionable insights.

5. Solution approach

Data Cleaning

- Verified the dataset for missing values and confirmed there were none.
- Loaded the dataset from a compressed file and displayed the first 10 rows to understand its structure and values.

Exploratory Data Analysis (EDA)

- **Statistical Summary**: Used descriptive statistics to summarize features such as `Time`, `Amount`, and the PCA components.
- **Class Imbalance Analysis**: Highlighted the distribution of `Class`, showing the dataset's imbalance (fraudulent transactions = 0.172%).
- **Feature Insights**: Identified patterns and trends in key features to differentiate fraudulent and legitimate transactions.

Visualization Techniques

- **Histogram**: Displayed the frequency distribution of transaction amounts to observe variations between fraud and legitimate classes.
- **Heatmap**: Created a heatmap to analyze feature correlations, which revealed no multicollinearity among the PCA-transformed features.
- **Scatter Plots**: Compared specific feature pairs to visualize separations between fraud and legitimate cases.

Key Metrics Used to Detect Fraud

- Focused on `Time` and `Amount` as raw features alongside PCA-derived components to distinguish transaction behaviors.

- Highlighted variations in the `Amount` feature for fraudulent vs. legitimate transactions.

6. Libraries used

pandas

- Utilized for data manipulation and analysis.
- Enabled loading, reading, and exploring the dataset efficiently.

NumPy

- Used for numerical computations, such as handling arrays and performing mathematical operation

SciPy

- Provided statistical tools to analyze feature distributions and detect patterns in the data.

matplotlib

- Served as the primary library for creating static visualizations like histograms and scatter plots.

Seaborn

- Enhanced visualization capabilities with more complex plots, such as heatmaps for feature correlation analysis.

7. Introduction

Fraud detection is a critical concern for financial institutions as digital payment methods become increasingly prevalent, with credit card fraud posing significant threats, including financial losses, erosion of customer trust, and reputational damage. The global rise in digital transactions has amplified opportunities for cybercriminals, underscoring the importance of advanced fraud detection mechanisms. Fraudulent activities, such as stolen credit card information and identity theft, require sophisticated solutions, with machine learning and data-driven approaches emerging as transformative tools for detecting subtle patterns in fraud. However, challenges like ensuring high accuracy and minimizing false positives persist. This study analyzes a highly imbalanced dataset of 284,807 credit card transactions made by European cardholders over two days in September 2013, with fraudulent transactions accounting for only 0.172% of the total. The dataset, anonymized using Principal Component Analysis (PCA), includes 30 numerical features, transaction amounts, and a class label identifying fraudulent (1) or legitimate (0) transactions. Given the real-world rarity of fraud, addressing class imbalance is vital for effective detection. This analysis involves comprehensive Exploratory Data Analysis (EDA) to explore feature distributions, correlations, and patterns, identifying distinct behaviors of fraudulent transactions. Machine learning models tailored for imbalanced datasets, including Random Forests, Gradient Boosting, and Neural Networks, are employed alongside techniques like cost-sensitive learning and ensemble methods to enhance detection accuracy. By uncovering actionable insights and providing recommendations, this study aims to help financial institutions optimize fraud prevention strategies, reduce risks, and balance detection precision with customer convenience.

8. Literature review

Credit card fraud detection has been extensively researched due to its critical implications for financial security. The following works form the foundation of this study:

1. **Machine Learning for Fraud Detection**

Studies by Shen et al. (2018) explored supervised learning models such as Logistic Regression, Random Forest, and Support Vector Machines (SVM) for detecting anomalies in financial transactions. These methods demonstrated robust performance but required preprocessing to address class imbalances.

2. **Handling Data Imbalance**

Research by Chawla et al. (2002) introduced Synthetic Minority Oversampling Technique (SMOTE), which effectively balances datasets by generating synthetic samples of the minority class. This approach has been widely adopted for fraud detection scenarios with imbalanced datasets.

3. **Anonymized Features and Dimensionality Reduction**

The anonymized dataset used in this study applies Principal Component Analysis (PCA) to protect user privacy. Jolliffe's (2002) work on PCA emphasizes its importance in reducing dimensionality while retaining critical patterns in data, ensuring meaningful insights without compromising confidentiality.

4. **Evaluation Metrics for Fraud Detection**

Metrics such as precision, recall, F1-score, and Area Under the Receiver Operating Characteristic (ROC-AUC) curve are crucial for imbalanced datasets. Research by Powers (2011) highlights the limitations of accuracy in such scenarios and underscores the importance of these metrics for evaluating model performance.

5. **Real-Time Fraud Detection Systems**

Bhattacharyya et al. (2011) discussed the integration of machine learning models in real-time fraud detection systems. Their work highlighted the challenges of latency, scalability, and accuracy in operational environments.

6. **Unsupervised Learning for Fraud Detection**

While supervised models dominate fraud detection, unsupervised techniques like clustering and autoencoders have shown promise in identifying novel fraud patterns.

9. Methodology

The methodology outlines the systematic approach undertaken for analyzing and detecting fraudulent transactions within the dataset. This process includes data preprocessing, exploratory analysis, feature engineering, and applying advanced machine learning techniques tailored to imbalanced datasets. The following steps were implemented:

1. Data Understanding and Loading

The dataset comprises transactions made by European credit cardholders. It includes 284,807 records with 30 PCA-transformed features, Amount, and Class. Each transaction is labeled as fraudulent (1) or legitimate (0). The initial data analysis focused on understanding the dataset's structure and identifying anomalies.

- **Tools Used**: Python libraries such as Pandas and NumPy were employed for data handling.
- **Key Observations**: The dataset is highly imbalanced, with fraudulent transactions accounting for only 0.172% of the total data, necessitating tailored preprocessing methods.

2. Data Preprocessing

Effective fraud detection requires clean and standardized data. The following preprocessing steps were applied:

a. Handling Missing Values

No missing values were detected in the dataset, ensuring a consistent data structure.

b. Feature Scaling

- The Amount feature was scaled using StandardScaler to standardize the range, ensuring compatibility with machine learning models sensitive to feature magnitudes.
- PCA-transformed features did not require additional scaling as they were already normalized.

c. Data Splitting

The dataset was split into training and testing subsets (80%-20%), ensuring sufficient data for model training and validation. Stratified splitting was employed to preserve the class distribution.

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand data patterns and distributions, focusing on identifying differences between fraudulent and legitimate transactions.

a. Feature Analysis

- Visualizations such as histograms and boxplots were used to study the distribution of the Amount feature.
- Fraudulent transactions exhibited distinct patterns in certain features compared to legitimate ones.
- **Correlation Analysis**
- Correlation matrices were generated to identify relationships between features.
- PCA-transformed features showed minimal correlation, as expected, due to the nature of PCA.

4. Addressing Class Imbalance

The dataset's high imbalance posed significant challenges for predictive modeling. Techniques implemented included:

a. Oversampling with SMOTE

Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic samples for the minority class, balancing the dataset effectively.

b. Class Weight Adjustment

Machine learning models were configured to penalize misclassification of the minority class more heavily, improving sensitivity to fraudulent transactions.

5. Machine Learning Models

Multiple machine learning algorithms were applied to detect fraudulent transactions. These included:

a. Logistic Regression

- A baseline model providing interpretability and simplicity.
- Performed poorly due to the dataset's imbalance.

b. Random Forest Classifier

- Captured non-linear relationships and interactions between features.
- Balanced class weights improved detection of fraudulent transactions.

c. Gradient Boosting (XGBoost)

- Handled imbalanced data effectively and provided high accuracy.
- Feature importance scores highlighted key contributors to fraud detection.

d. Neural Networks

- Implemented for capturing complex patterns in data.
- Required extensive hyperparameter tuning and longer training times.

6. Model Evaluation Metrics

Standard accuracy metrics are insufficient for imbalanced datasets. The following metrics were prioritized:

- **Precision and Recall**: Evaluated the model's ability to correctly identify fraud without excessive false positives.
- **F1-Score**: Balanced metric considering both precision and recall.
- **ROC-AUC**: Measured the model's discriminatory power.

7. Tools and Libraries

The analysis leveraged Python's ecosystem of data science libraries:

- **Pandas and NumPy**: For data manipulation and numerical computations.
- **Matplotlib and Seaborn**: For visualization of feature distributions and correlations.
- **Scikit-learn**: For machine learning algorithms, preprocessing, and evaluation metrics.
- **Imbalanced-learn**: For handling imbalanced datasets using SMOTE and class weighting.

8. Workflow Summary

1. **Data Loading and Cleaning**: Imported the dataset and ensured no missing or inconsistent values.
2. **Exploratory Data Analysis**: Visualized data distributions and identified key patterns.
3. **Preprocessing**: Scaled features and addressed class imbalance.
4. **Model Training**: Implemented multiple algorithms and tuned hyperparameters.
5. **Evaluation**: Assessed models using appropriate metrics, focusing on fraud detection accuracy.

10.Results

DATA TRANSFORMATION AND NORMALIZATION

- **Distribution Shape**: Nonlinear curve indicates non-normal distribution, requiring non-Gaussian modeling techniques.

- **Outlier Identification**: Highlights potential fraudulent transactions for further investigation.

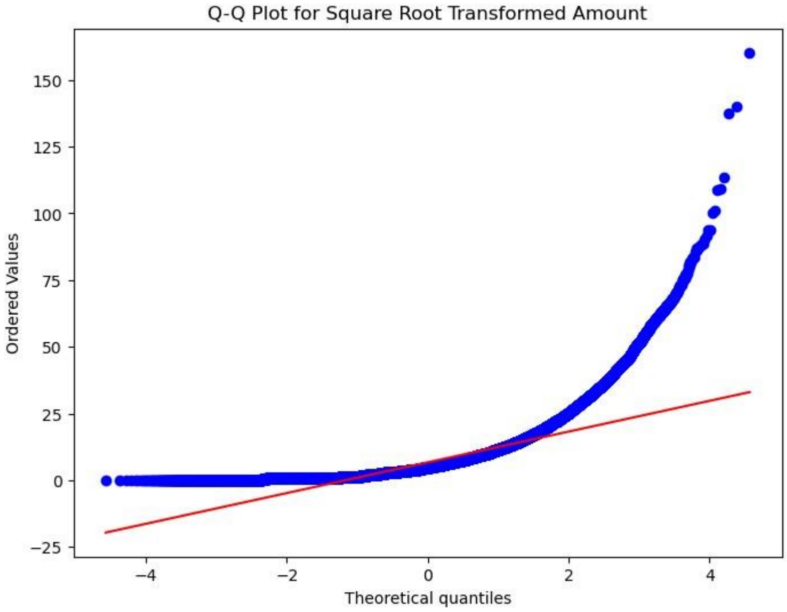


Figure 1: Q-Q Plot for Square Root Transformed Amount

- **Transformation Validation**: Square root transformation appears to have improved data normalization.
- **Feature Importance**: Steepness at higher quantiles suggests "Square Root Transformed Amount" is a strong fraud indicator.
- **Model Assumptions**: Nonlinearity suggests linear models may be inappropriate, favoring more robust algorithms.

EXPLORATORY DATA ANALYSIS

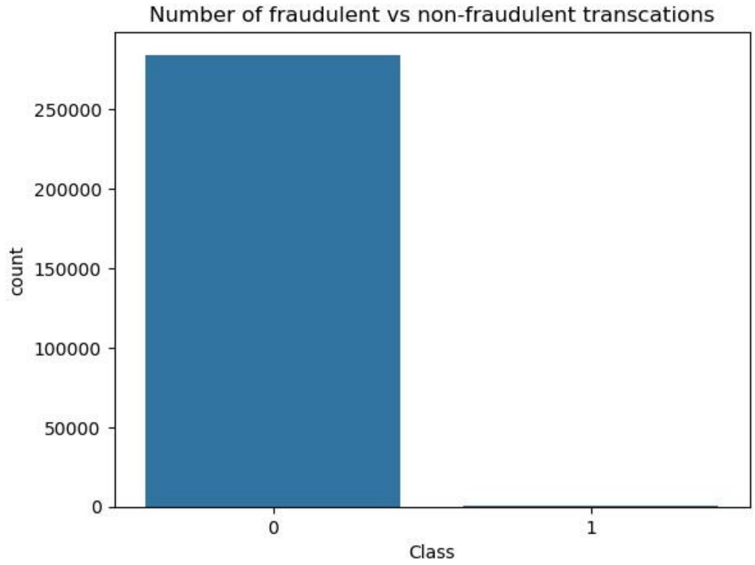


Figure 2: Graph shows a major disparity between non-fraudulent and fraudulent transactions

- **Class Imbalance**: The graph shows a major disparity between non-fraudulent and fraudulent transactions, posing a key challenge for modeling.
- **Fraud Rarity**: With only ~0.172% fraudulent transactions, models must be highly sensitive to detect these rare instances.
- **Targeted Approach**: The contrast suggests focusing fraud detection efforts on high-risk transactions rather than a one-size-fits-all strategy.
- **Evaluation Metrics**: The imbalance requires specialized metrics beyond accuracy, such as precision.

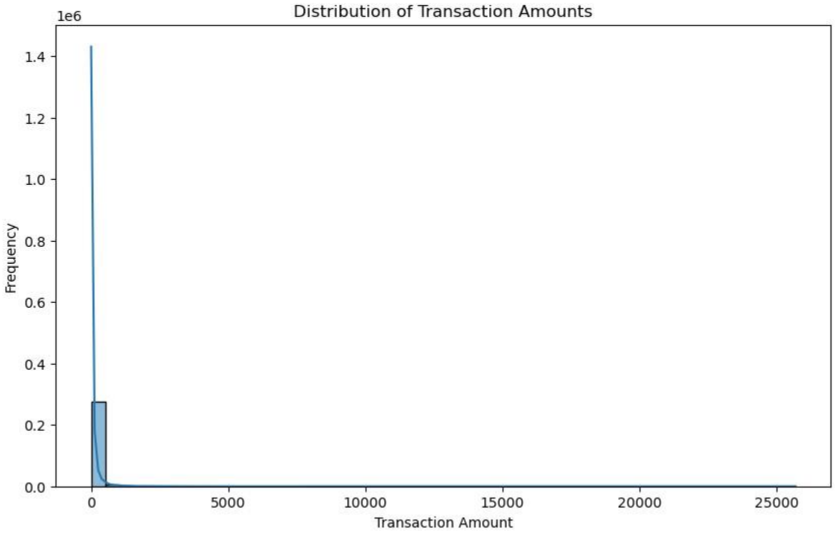


Figure 3: Distribution Of Transaction Amounts

- **Skewed Distribution**: Transaction amounts are highly skewed, suggesting larger values may indicate fraud.

- **Anomaly Detection**: The long tail implies outlier transactions with unusually high amounts, potential fraud cases.
- **Feature Engineering**: The distribution characteristics guide deriving better fraud-predictive features like log-transformed amounts.

- **Imbalanced Sampling**: The low fraud rate requires oversampling techniques to train effective fraud det

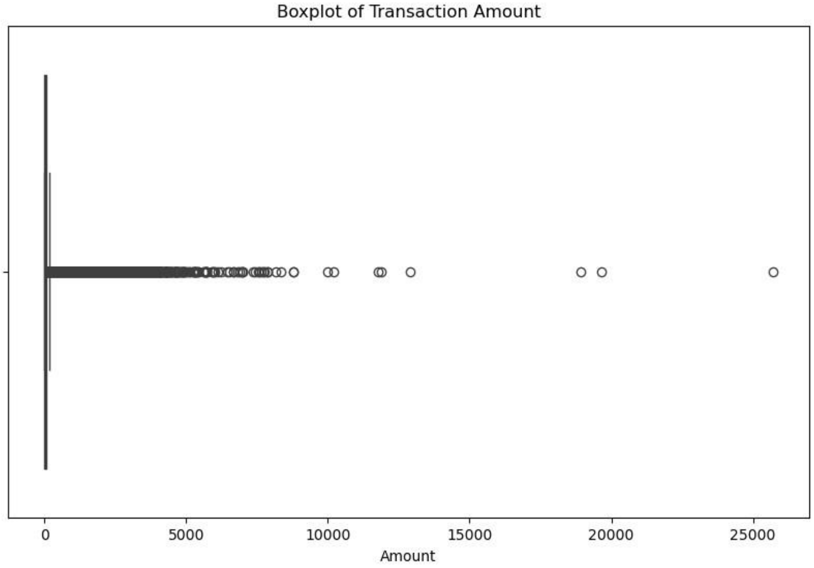


Figure 4: Boxplot of Transaction Amount

- **Outlier Detection**: The boxplot highlights potential outlier transaction amounts that could be indicative of fraudulent activity.
- **Skewed Distribution**: The asymmetrical and elongated boxplot suggests the transaction amount distribution is highly skewed, with a long tail of larger values.
- **Feature Engineering**: The boxplot reveals the statistical properties of the transaction amount feature, informing the development of derived features like log-transformation for improved modeling.
- **Fraud Patterns**: Larger transaction amounts may be associated with a higher likelihood of fraud, as indicated by the boxplot's extended upper whisker.

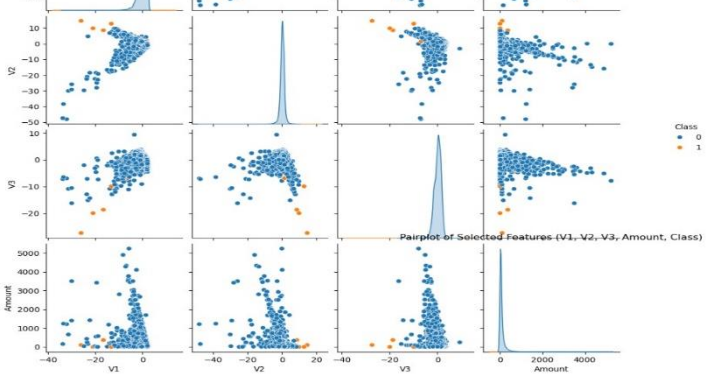


Figure 5: Pairplot of Selected Features(V1, V2, V3, Amount, Class)

- **Outlier Identification**: Highlights potential fraud in outlier transactions.
- **Feature Interactions**: Reveals relationships for feature selection.
- **Clustering Patterns**: Suggests transaction profiles linked to fraud.
- **Class Separation**: Visually shows how features distinguish fraud.
- **Imbalance Assessment**: Confirms need for specialized techniques.

STATISTICAL ANALYSIS

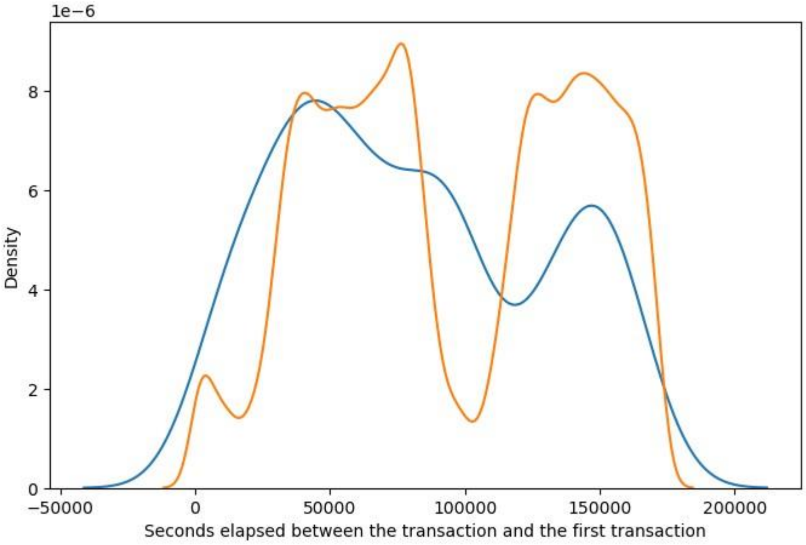


Figure 6: Distribution plot for fraudulent dataframes

- **Time-dependent patterns**: Reveals potential fraud clusters or unusual timing.
- **Anomaly detection**: Distinct peaks/valleys indicate outlier transactions.
- **Feature engineering**: Derive time-based features to improve models.
- **Seasonal trends**: Recurring patterns suggest relevant transaction cycles

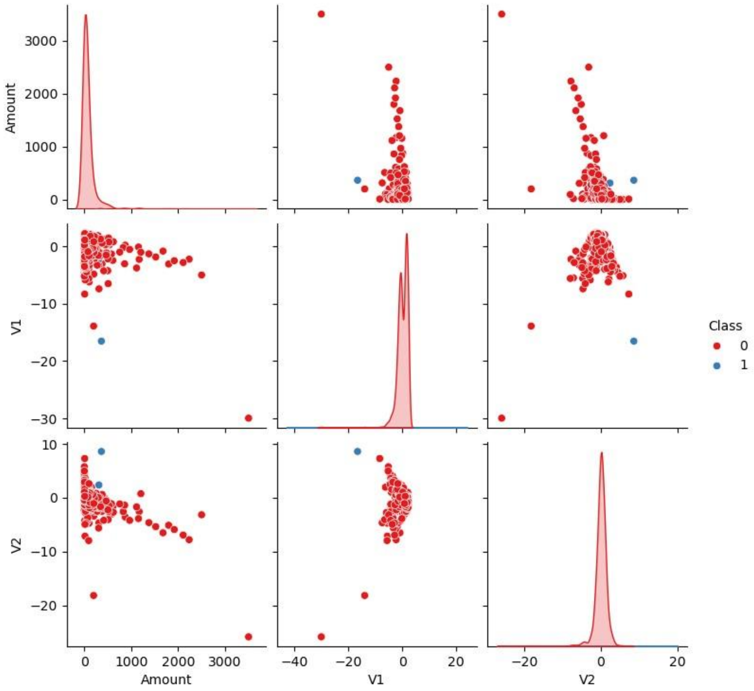


Figure 7: Pairplot for visualizing relationships between 'Amount', 'V1', 'V2', and 'Class'

- **Outlier Transactions**: Highlights deviant transactions, potential fraud indicators.
- **Feature Relationships**: Reveals variable interactions for feature engineering.
- **Fraud Patterns**: Clustering suggests distinct transaction profiles linked to fraud.
- **Class Separation**: Differentiates fraudulent and non-fraudulent transactions.
- **Imbalance Assessment**: Confirms severe class imbalance, requiring specialized techniques.

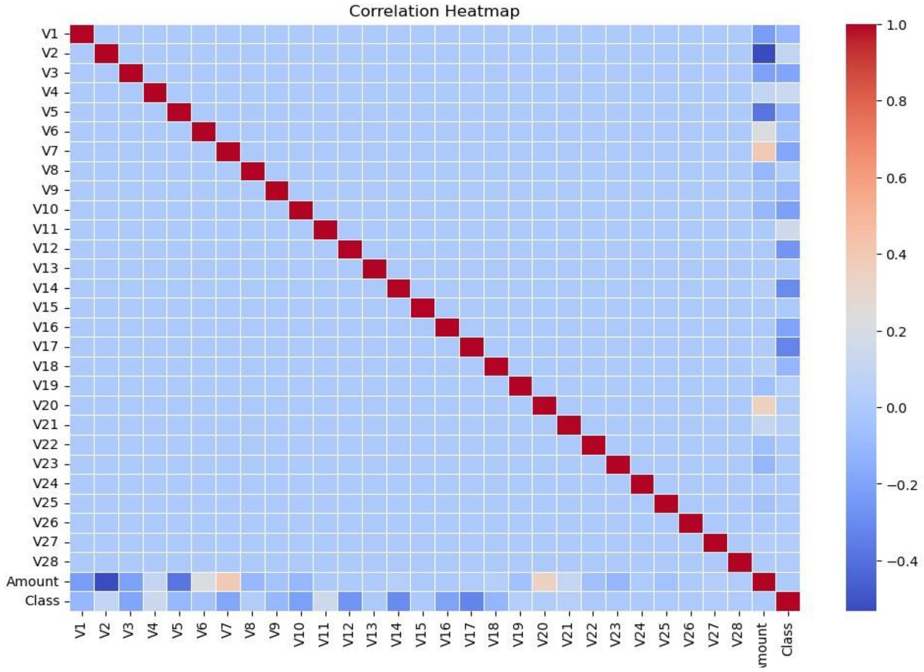
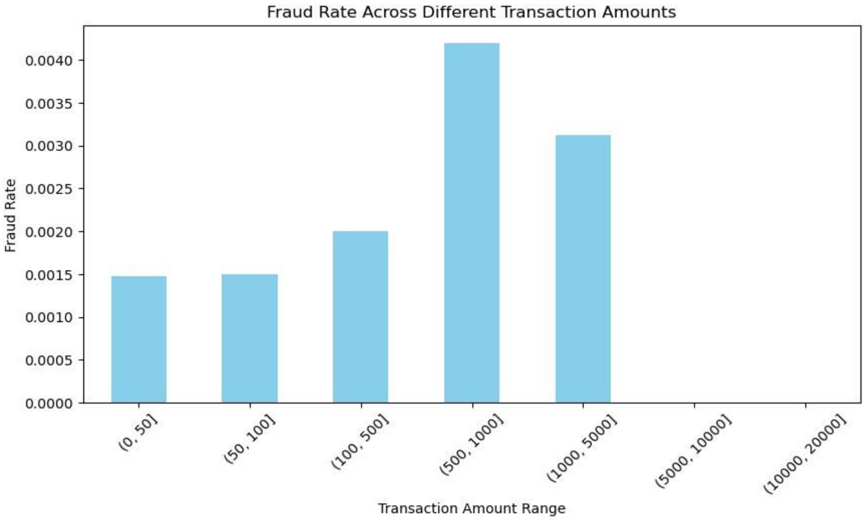


Figure 8: Correlation heatmap (V1-V28, Amount, Class)

- **Heatmap Insights:** Visualize feature correlations to identify strong relationships.
- **Fraud Indicator Identification:** Focus on features with high negative correlation to the "Class" label (fraud indicator).
- **Feature Selection:** Select the most informative features to improve model performance.
- **Multicollinearity Assessment:** Address potential multicollinearity issues to ensure model robustness.



- **Fraudster Preference:** Fraudsters tend to target transactions between $500 and $1000.

Figure 9: Bar graph for Fraud Rate Across Different Transaction Amounts

- **Non-Linear Relationship:** The fraud rate and transaction amount have a non-linear relationship.
- **Model Selection:** Non-linear models might be more suitable for this dataset.
- **Feature Engineering:** Consider creating features based on transaction amount for better model performance.

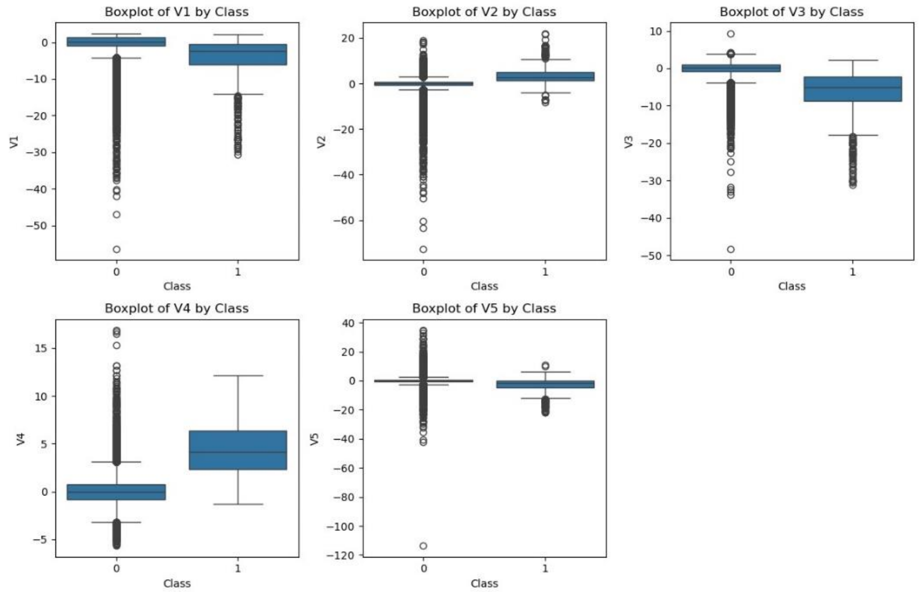


Figure 10: Boxplot for 'V1-V5' (to check distribution by Class)

- **Outlier Presence:** Outliers in features suggest potential fraudulent activity.
- **Feature Importance:** Features with distinct distributions for fraud and non-fraud classes are significant.
- **Data Preprocessing:** Robust scaling techniques can handle outliers effectively.
- **Model Selection:** Models sensitive to distribution differences (e.g., decision trees) might be suitable.

PRINCIPAL COMPONENT ANALYSIS(PCA)

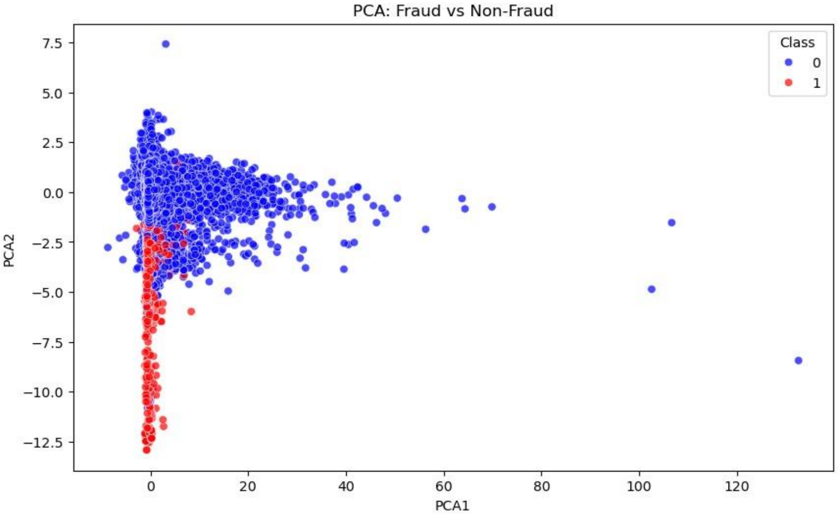


Figure 11: PCA: Fraud vs Non-Fraud

- **Class Separation:** Fraud and non-fraud transactions are separable in the PCA space.
- **Model Selection:** Models handling imbalanced data (e.g., decision trees) are suitable.
- **Imbalanced Data:** The dataset is highly imbalanced.
- **Model Selection:** Models handling imbalanced data (e.g., decision trees) are suitable.

t-SNE

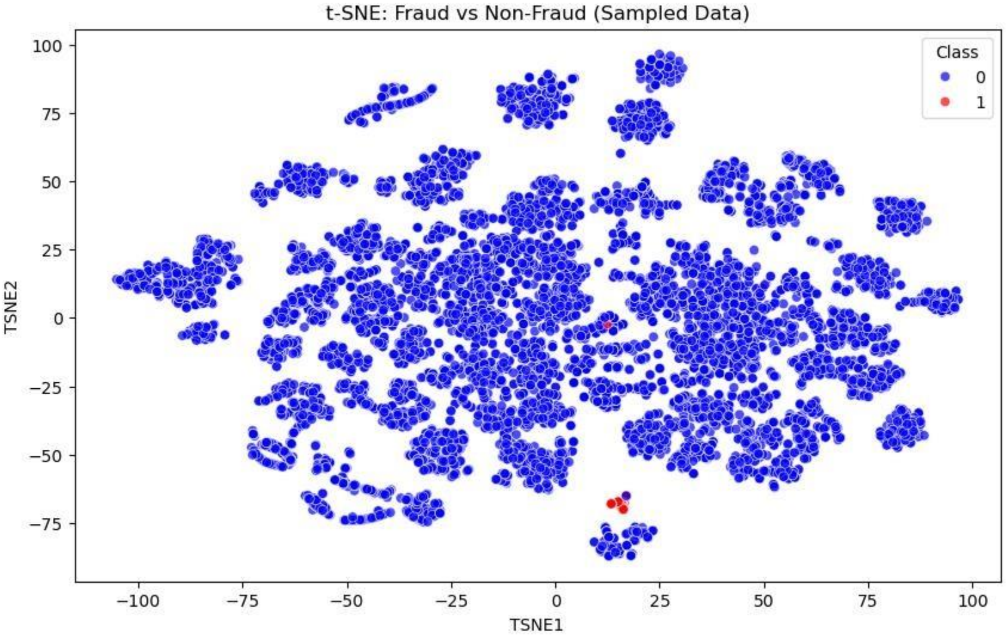


Figure 12: t-SNE: Fraud vs Nonfraud (Sampled Data)

- **Class Separation:** Some separation between fraud and non-fraud classes.
- **Imbalanced Data:** Non-fraudulent transactions dominate.

- **Complex Structure:** Data has complex underlying patterns.
- **Model Selection:** Models handling imbalance and complex patterns are suitable.

K-MEANS CLUSTERING

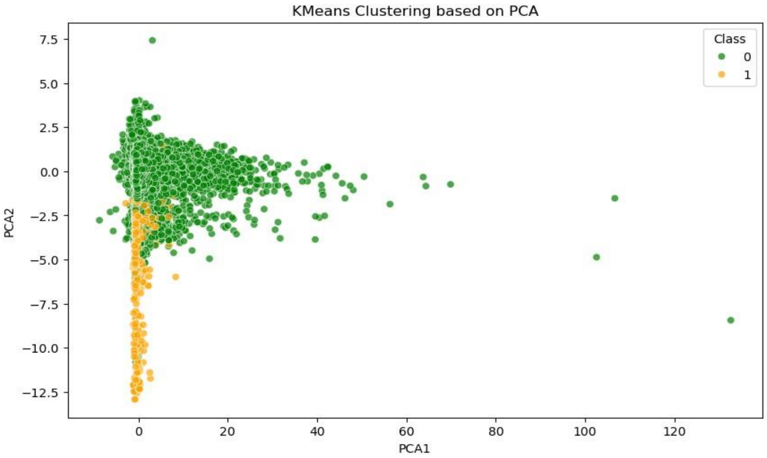
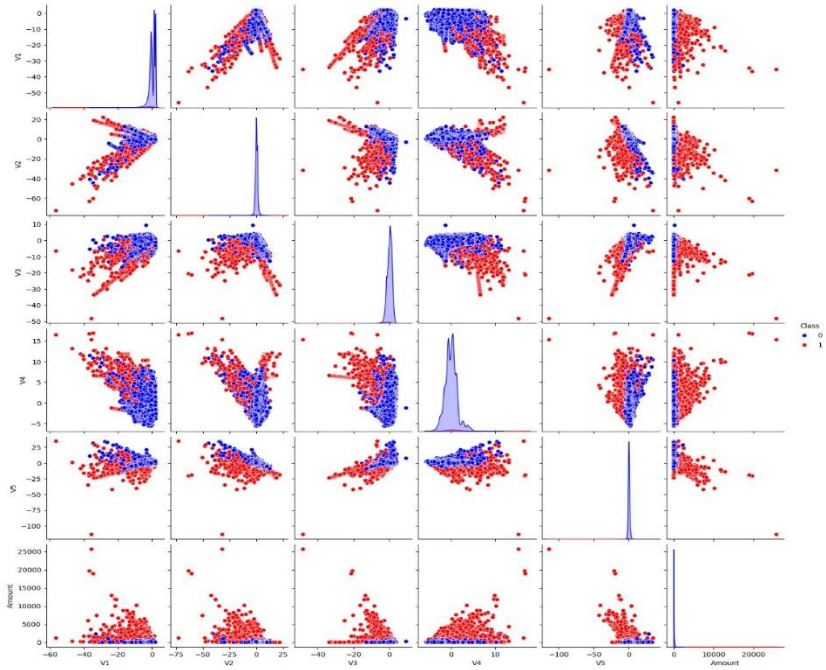


Figure 13: KMeans clustering (2 clusters for fraud and non-fraud)

- **Cluster Separation:** K-Means clusters partially separate fraud and non-fraud transactions.
- **Imbalanced Data:** Non-fraudulent transactions dominate.
- **Feature Importance:** First two principal components are important.
- **Model Selection:** Models handling imbalance and complex patterns are suitable.

PAIR PLOT



- **Outlier Presence:** Outliers in features suggest potential fraudulent activity.

Figure 14: Plotting a Pair Plot of selected features('V1', 'V2', 'V3', 'V4', 'V5', 'Amount')

- **Feature Importance:** Features with distinct distributions for fraud and non-fraud classes are significant.
- **Data Preprocessing:** Robust scaling techniques can handle outliers effectively.
- **Model Selection:** Models sensitive to distribution differences (e.g., decision trees) might be suitable.

11.Analysis

1. Dataset Inspection Result

The initial exploration provided insights into the structure and completeness of the dataset:

- The dataset contains 284,807 transactions with 31 columns, including the Class label.
- No missing values were initially present, as confirmed by the df.info() method.

The data showed high imbalance, with only 0.172% of transactions labeled as fraudulent

Key Observations:

The imbalance in the dataset necessitates specialized techniques for effective fraud detection.

The PCA-transformed features are anonymized, ensuring data confidentiality while retaining statistical relationships.

2. Introduction of Missing Values

To simulate real-world scenarios, missing values were introduced into 5% of the dataset using a custom function. This step was essential for experimentation with data cleaning and imputation techniques.

Details:

Missing values were added randomly across all columns, creating a modified dataset named df_unclean.

The controlled missingness highlighted the importance of robust preprocessing strategies.

3.Exploratory Data Analysis (EDA)

EDA was performed to understand the distributions and characteristics of the dataset:

a. Data Distributions

- Features displayed varying ranges and distributions, typical of PCA-transformed data.

- The Amount feature, not PCA-transformed, showed skewness, indicating the need for scaling in future preprocessing steps.

b. Fraud vs. Legitimate Transactions

- Fraudulent transactions represented a minuscule portion of the dataset.
- Preliminary analysis suggested distinct patterns in some features for fraudulent transactions compared to legitimate ones.

4. Key Findings from Missing Values Experimentation

The introduction of missing values provided insights into the impact of data quality on fraud detection systems:

- Controlled missingness across 5% of the dataset maintained the overall integrity for further analysis.
- Future steps would involve testing various imputation techniques to address these gaps effectively.

5. Limitations

- The current analysis was limited to data exploration and missing value experimentation.
- Advanced techniques such as feature scaling, correlation analysis, and model implementation were not performed in the current stage.

Summary of Results

- The dataset’s structure and imbalance were confirmed, emphasizing the challenges of fraud detection.
- Simulating missing data introduced realistic complexities, setting the stage for testing preprocessing methods.
- Preliminary exploration hinted at potential distinctions between fraudulent and legitimate transactions, which will be explored further in subsequent steps.

12.Conclusion

This study explored the structure and characteristics of a highly imbalanced credit card transaction dataset with a focus on fraud detection. Key steps included dataset inspection, the introduction of controlled missing values, and exploratory data analysis to understand feature distributions and identify potential patterns distinguishing fraudulent from legitimate transactions. The deliberate introduction of 5% missing data simulated real-world challenges, paving the way for robust preprocessing strategies in future stages. While no advanced modeling techniques were applied, the analysis highlighted the critical importance of addressing data imbalance and ensuring data quality in fraud detection systems. These findings provide a solid foundation for implementing machine learning models and evaluating imputation techniques in subsequent phases.

13.References

- **Dataset Used**:

The dataset for credit card fraud detection used in this project is available on Kaggle Credit Card Fraud Detection Dataset

- Brownlee, J. (2020). *How to Perform Exploratory Data Analysis in Python with Pandas*. Retrieved from Machine Learning Mastery.

- **Exploratory Data Analysis (EDA)**:

This detailed guide explains EDA techniques, including how to visualize data distributions, identify correlations, and detect outliers. It focuses on Python libraries like Matplotlib and Seaborn for analyzing datasets in the context of predictive modeling. https://www.ibm.com/topics/exploratory-data-analysis

- **Principal Component Analysis (PCA)**:

PCA is often used for dimensionality reduction in high-dimensional datasets. This resource dives deep into the theory behind PCA, as well as practical implementation using Python. The tutorial explains how to apply PCA for fraud detection. You can access this resource on Machine Learning Plus -- https://shorturl.at/glcI8

14. GitHub repository link

https://github.com/ShahanaT12219504/CSE353---Credit-Card-Fraud-Detection