Enhancing Credit Card Fraud Detection

Literature Review

Submitted By: **Gaurang Dadu**

Student Number: **501262373**

Date: **October 18th, 2023**

Supervisor's Name: **Dr. Ceni Babaoglu**

Ryerson University

Table of Contents

1. ABSTRACT

Credit Card Fraud Detection is a critical concern within the financial sector, demanding advanced techniques for precise identification and prevention. This research paper delves into the realm of Big Data and Predictive Analysis, aiming to develop and compare three distinct machine learning models - Random Forest, Decision Tree, and Logistic Regression - for the purpose of accurate credit card fraud detection. Leveraging a publicly available dataset from Kaggle [1], this study employs these diverse models to create algorithms with the highest predictive accuracy.

The primary objective of this research is to ascertain which of these three machine learning models outperforms the others in predicting credit card fraud. This paper also addresses fundamental research questions, including an exploration of feature importance and the application of feature engineering techniques to optimize model performance. Additionally, it investigates the relationship between different transaction types and their correlation with fraudulent activities, uncovering latent patterns and trends.

To conduct this research, the study relies on popular data analysis libraries in Python and R, including Pandas, NumPy, Scikit-Learn, dplyr, ggplot2 and etc. These libraries serve as invaluable tools for data preprocessing, model development, and performance evaluation, offering extensive capabilities for analyzing the credit card fraud dataset.

The paper systematically evaluates the three models to determine their respective performance levels in credit card fraud detection. Each of these models brings unique strengths and characteristics to the task, providing a comprehensive analysis of their capabilities.

In terms of performance evaluation, this research paper employs robust methodologies, such as Confusion Matrix and F1 Score, to assess the effectiveness of the three machine learning models. These metrics offer a holistic view of the precision, recall, and overall performance of these models in detecting fraudulent transactions.

In conclusion, this research paper provides valuable insights into credit card fraud detection by comparing and evaluating three distinct machine learning models - Random Forest, Decision Tree, and Logistic Regression. The findings will assist financial institutions in implementing robust fraud prevention measures and contribute to the broader field of predictive analytics and data-driven decision-making.

Keywords**:** *Credit Card Fraud Detection, Machine Learning Models, Random Forest, Decision Tree, Logistic Regression, Feature Importance, Feature Engineering, Transaction Types, Performance evaluation, Predictive Analysis.*

2. LITERATURE REVIEW

2.1. Introduction:

In the complex landscape of financial transactions, credit card fraud looms as a persistent and multifaceted challenge. As our world increasingly embraces digital payments and online commerce, the need for robust fraud detection systems becomes more pressing. In response to this demand, researchers and data scientists have committed substantial resources to enhance the precision and effectiveness of credit card fraud detection mechanisms. This comprehensive literature review is a deep dive into this extensive body of research, which serves as the foundation for my ongoing investigation.

The realm of credit card fraud detection research is dynamic, characterized by a relentless pursuit of strategies to fortify the predictive powers of machine learning models. This quest is driven by the ever-evolving tactics employed by those who seek to defraud financial systems. This literature review aims to provide an encompassing overview of existing studies, methodologies, and advancements, addressing the evolving challenges that underlie the detection of credit card fraud.

The exploration will scrutinize an array of studies, delving into the technical intricacies and nuances inherent to each. I will shine a spotlight on the algorithms used, the techniques for selecting valuable features, and the metrics employed to evaluate model performance. My

objective is to gain insights into the technical expertise harnessed by the data science community to build models with exceptional accuracy in identifying fraudulent activities.

Feature engineering will emerge as a central theme within this review. Feature engineering is the art of meticulously crafting and selecting features from raw data. This process serves as the linchpin for enhancing the resilience of predictive models. I will investigate how researchers have ingeniously constructed features that capture subtle yet critical patterns indicative of fraudulent transactions.

The connection between various transaction types and their relationship to fraudulent activities will be a focal point of our exploration. In this context, I will dissect how transaction-specific attributes influence the predictive capabilities of fraud detection models. I will explore the technical dimensions of feature extraction and the creation of transaction-specific variables, shedding light on the importance of these aspects.

Moreover, this review will identify gaps in the existing literature that warrant further exploration. This identification forms the bedrock for the research, enabling me to address these gaps and expand upon the technical knowledge and insights gained from previous studies. My ultimate goal is to develop and meticulously scrutinize three distinct machine learning models—Random Forest, Decision Tree, and Logistic Regression. These models hold the potential to elevate credit card fraud detection and prevention to a new level, underscoring our commitment to fortifying financial systems against this unrelenting challenge.

This literature review serves as a comprehensive prologue to my research, encapsulating the intricate technicalities, methodological rigor, and unresolved challenges that define the continually evolving field of credit card fraud detection. It establishes the foundation for my contribution to the sphere of predictive analytics and data-driven decision-making in the domain of credit card fraud detection, emphasizing the technical nuances that underpin this dynamic field.

2.2. Related Work:

In recent years, the financial sector has witnessed a surge in credit card fraud, compelling the need for sophisticated detection techniques. The study conducted by ***"Jonathan Kwaku Afriyie et al"***, in their paper titled ***"A supervised Machine Learning Algorithm for Detecting and Predicting Fraud in Credit Card Transactions"*** [2], explores the performance of three distinct machine learning models, namely logistic regression, random forest, and decision trees, in the realm of credit card fraud detection. Their comprehensive analysis reveals that random forest emerges as the most effective model, achieving a remarkable accuracy of 96% and an area under the curve (AUC) value of 98.9% in identifying fraudulent credit card transactions. This finding underscores the significance of employing advanced machine learning algorithms in tackling the growing menace of credit card fraud. Moreover, their investigation uncovers intriguing patterns, such as a higher incidence of fraud among credit card holders aged above 60 and a concentrated timeframe for fraudulent activities between 22:00GMT and 4:00GMT. The insights derived from this research are invaluable for shaping more robust fraud prevention measures and predictive analytics in the financial sector.

In the realm of credit card fraud detection, ***"K.R. Seeja and Masoumeh Zareapoor"***, in their paper titled, ***"FraudMiner"*** [3], present an innovative approach to address the challenges posed by highly imbalanced and anonymous credit card transaction datasets. The paper introduces an intelligent model that tackles class imbalance by identifying legal and fraud transaction patterns for each customer through frequent itemset mining. It employs a matching algorithm to determine whether incoming transactions align more closely with legal or fraudulent patterns, enabling effective fraud detection. Unlike traditional methods, the model does not favor any specific attributes, considering each attribute equally to derive patterns. The model's performance is evaluated using the UCSD Data Mining Contest 2009 Dataset, demonstrating impressive results with a high fraud detection rate, balanced classification rate, Matthew's correlation coefficient, and low false alarm rate compared to other state-of-the-art classifiers. This research is a significant contribution to the field, especially in handling anonymous and imbalanced credit card transaction data, offering a valuable perspective for future fraud detection systems.

The paper authored by ***"Swati Warghade et al"***, titled ***"Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm"*** [4], delves into the intricacies of credit card fraud detection, presenting a pioneering approach to mitigate the mounting threat of fraudulent credit card transactions. In an era where credit cards have become the predominant means of conducting both online and offline transactions, the specter of credit card fraud looms large, inflicting significant financial losses on the industry. This research paper centers on the formidable challenge posed by highly imbalanced credit card fraud datasets, where the prevalence of legitimate transactions vastly outnumbers fraudulent ones. To confront this issue head-on, the paper explores an array of machine learning techniques and employs diverse metrics to assess various classifiers. Its core objective is to enhance the efficacy of fraud

detection while minimizing the misclassification of legitimate transactions as fraudulent. Notably, the paper highlights the critical role of **Synthetic Minority Oversampling Technique (SMOTE)** in addressing highly imbalanced datasets, providing a valuable solution to optimize fraud detection strategies and safeguard the financial sector against the evolving landscape of financial fraud.

In the landscape of fraud detection in electronic transactions, the research conducted by ***"Rafael Lima and Adriano Pereira"***, titled ***"Feature Selection Approaches to Fraud Detection in e-Payment Systems"*** [5], offers a comprehensive exploration into the intricacies of feature selection as a critical aspect of fraud identification. With the surge in e-commerce transactions and the accompanying rise in online fraud, the need for effective fraud detection strategies has become paramount. However, the inherent class imbalance between fraud and non-fraud transactions poses significant challenges for traditional feature selection techniques. This study meticulously examines the impact of class imbalance on feature selection, revealing that traditional methods may not be optimal for identifying anomalies, especially in the context of highly imbalanced datasets. To address these challenges, the research delves into the effectiveness of resampling methods before the feature selection step. Seven different resampling approaches, including a novel method developed by the authors, are rigorously evaluated. The research also culminates in the creation of a robust model for fraud detection, combining resampling strategies and classification techniques. The findings provide valuable insights into mitigating the impact of class imbalance and optimizing feature selection for enhanced fraud detection. This research, grounded in real-world data from a prominent Latin American electronic payment system, represents a significant contribution to the field of fraud detection in electronic transactions.

The research paper authored by ***"Bahnsen et al"***, titled ***"Feature Engineering Strategies for Credit Card Fraud Detection"*** [6], makes a significant contribution to the field of credit card fraud detection by focusing on feature engineering to enhance model accuracy. With the ever-increasing prevalence of credit card fraud leading to billions of Euros in financial losses annually, it has become imperative for financial institutions to continually improve their fraud detection systems. This study addresses the limitations of previous research, particularly the omission of the actual financial costs involved in fraud detection and the need for advanced feature engineering techniques. The paper introduces an innovative approach to create new features, emphasizing the extraction of periodic features using the von Mises distribution. Through rigorous experimentation with a real credit card fraud dataset provided by a major European card processing company, the research demonstrates the significant impact of these proposed features on fraud detection results. The outcomes reveal an average increase in savings of 13% when employing the periodic features. Furthermore, these findings are currently being employed to develop an advanced fraud detection system. This paper offers a comprehensive exploration of enhanced feature engineering techniques and financial cost-sensitive evaluation measures, making it an essential reference for those seeking to improve credit card fraud detection systems.

This comprehensive research paper authored by ***"Asma Cherif et al"***, titled ***"Credit Card Fraud Detection in the era of Disruptive Technologies: A Systematic Review"*** [7], offers a timely and thorough review of credit card fraud detection methodologies, focusing on the period from 2015 to 2021. In light of the evolving landscape of financial technology, including

contactless payments and digital commerce, the authors address the pressing need for robust fraud detection systems. They meticulously analyze 40 pertinent articles, categorizing them based on topics such as addressing the class imbalance problem, feature engineering, and the utilization of machine learning technologies, both traditional and deep learning. Notably, this study reveals that the adoption of deep learning in credit card fraud detection remains underexplored, highlighting the necessity for further research in this domain. As financial institutions increasingly rely on new technologies like big data analytics, large-scale machine learning, and cloud computing, this paper provides invaluable insights into ongoing research challenges and future directions. It serves as a valuable resource for academic and industrial researchers striving to enhance the efficacy of financial fraud detection systems and develop resilient solutions.

2.3. **Purpose of the Study**:

This research fits within the existing body of work in credit card fraud detection by building upon the foundations laid by previous studies. In this literature review, I have provided a comprehensive overview of various methodologies, approaches, and insights that have been developed in this field, highlighting both the challenges and opportunities in the domain of credit card fraud detection.

This research is worth conducting for several reasons:

1. **Evolution of Technology:** As the financial landscape continues to evolve with the adoption of innovative payment technologies and digital commerce, this study addresses the evolving challenges and the need for robust detection systems in this context.

2. **Feature Engineering:** This research emphasizes the importance of feature engineering, which is a critical aspect of building effective fraud detection models. By exploring feature engineering techniques, this work aims to enhance the accuracy of models, filling a critical gap identified in the literature.
3. **Deep Learning Exploration:** Despite the rapid advances in deep learning technologies, the adoption of these methods in credit card fraud detection remains relatively unexplored. This study points out this gap and the necessity for further research in the application of deep learning to enhance fraud detection.
4. **Real-World Relevance:** This research leverages real-world datasets and pragmatic metrics, which are more applicable to practical fraud detection scenarios. The application of machine learning models to publicly available datasets from Kaggle underscores the real-world applicability of my work.

In summary, this research is worth conducting because it builds upon prior work, addresses evolving challenges in the financial sector, explores advanced feature engineering techniques, emphasizes the potential of deep learning, and seeks to contribute to more effective fraud detection. This work has the potential to advance the field of credit card fraud detection and address the growing challenges posed by evolving technologies and fraudulent activities.

3. DESCRIPTIVE STATISTICS OF THE DATASET

The dataset I'm working with is called the **"Online Payments Fraud Detection Dataset."** I discovered this dataset on Kaggle [1], and it's quite sizable, weighing in at 495 MB. It contains a substantial 6,362,620 rows of data, with each row representing a different online transaction. There are 11 distinct pieces of information (features) recorded for each transaction.

I selected this dataset for several good reasons. First, it's large, which provides me with a wealth of transaction records to analyze. Second, it's structured, meaning the data is neatly organized, making it straightforward for me to study. This structured format simplifies my analysis of the transaction data, which is essential for my research on credit card fraud detection.

What makes this dataset particularly noteworthy is that it has been used by many researchers before me. It's almost like a reference point because others have explored it as well. The dataset's history of use makes it a valuable resource for me to compare and refer to in my research.

Additionally, it's beneficial that this dataset is available on Kaggle, a platform where data scientists and researchers share their work. This aligns with the principles of open data, which encourage transparency, collaboration, and making research findings easy to check and verify. These principles are important in my research journey.

3.1. Data Variables Description

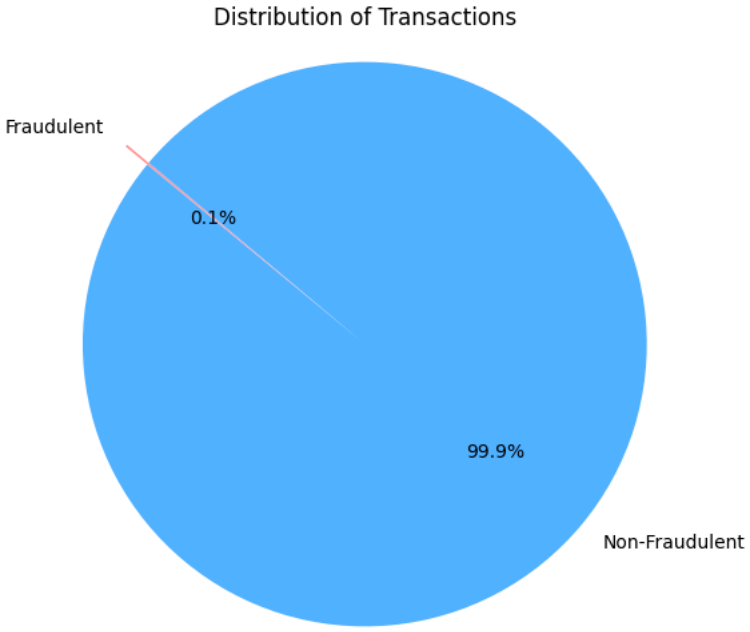
The dataset comprises several essential features that provide insight into these transactions. Here's a closer look at what each feature entails:

- **step:** This feature represents a unit of time in the dataset. Specifically, one step equates to one hour of time, allowing us to analyze the temporal aspects of each transaction.
- **type:** The "type" feature categorizes the online transactions into different types. It classifies whether a transaction involves a payment, transfer, withdrawal, or another type. Understanding the type of transaction is crucial for identifying any unusual patterns associated with fraudulent activities.
- **amount:** The "amount" feature indicates the specific monetary value of each transaction. This information is fundamental for assessing transaction patterns and detecting outliers.
- **nameOrig:** This feature identifies the customer initiating the transaction. Each customer is linked to their respective transactions, enabling us to track individual transaction behaviors.
- **oldbalanceOrg:** The "oldbalanceOrg" feature denotes the customer's account balance before a transaction takes place.
- **newbalanceOrig:** After a transaction, the customer's account balance is updated and recorded in the "newbalanceOrig" feature.

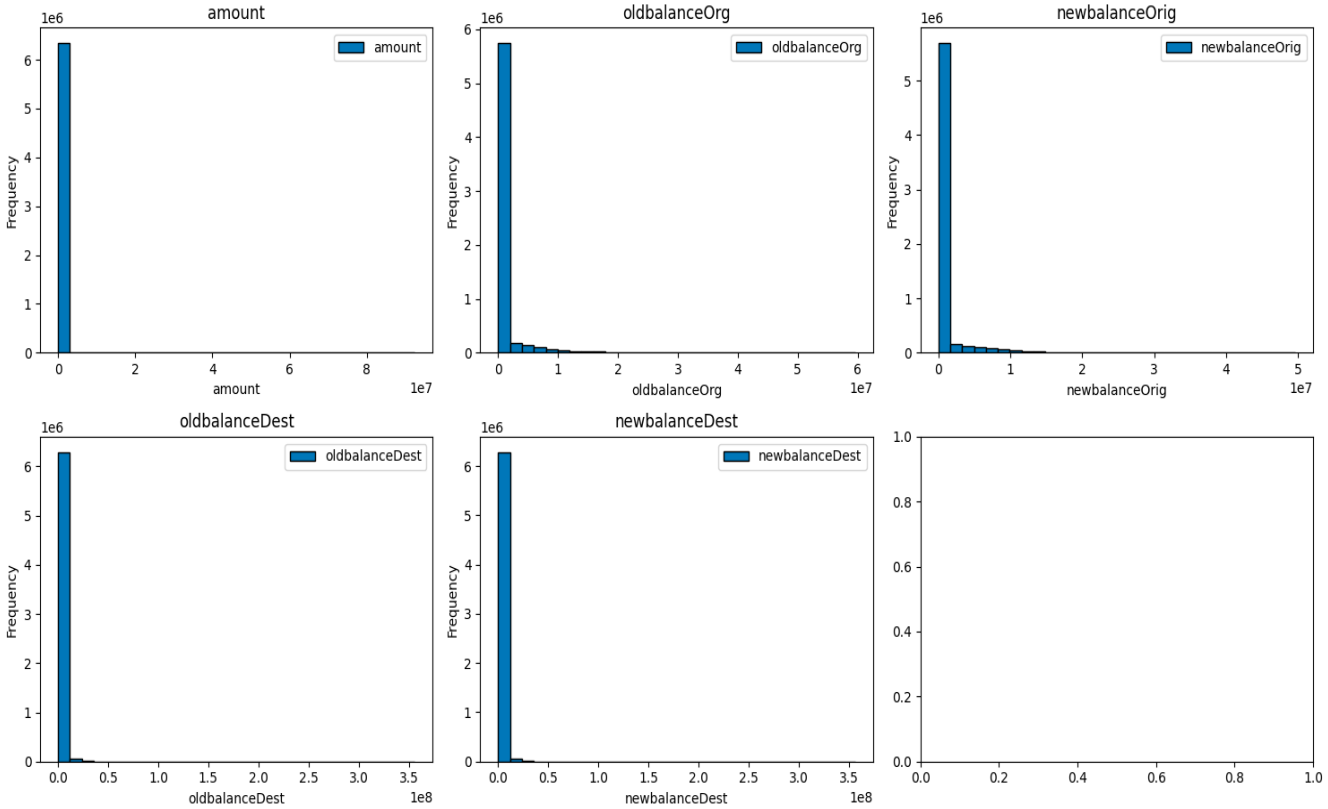
- **nameDest:** In the case of a transaction, there is typically a recipient. The "nameDest" feature specifies the recipient's information, allowing us to trace both ends of the transaction.
- **oldbalanceDest:** The "oldbalanceDest" feature signifies the recipient's initial account balance before the transaction is executed.
- **newbalanceDest:** Following the transaction, the recipient's account balance is adjusted and documented in the "newbalanceDest" feature.
- **isFraud:** The most critical feature of all, "isFraud," is a binary indicator that reveals whether a transaction is fraudulent or not. This feature is central to my research, as it serves as the target variable for credit card fraud detection.
- **isFlaggedFraud**: The "isFlaggedFraud" feature, in this context, serves as a binary indicator that determines whether a transaction flagged as potentially fraudulent by a specific algorithm is indeed marked as fraud. It plays a crucial role in assessing the accuracy and effectiveness of the algorithm's fraud detection capabilities, particularly in correctly identifying transactions that meet predefined criteria for being flagged as fraudulent. This feature contributes to the validation and fine-tuning of the fraud detection algorithm's performance.

3.2. Data Visualization

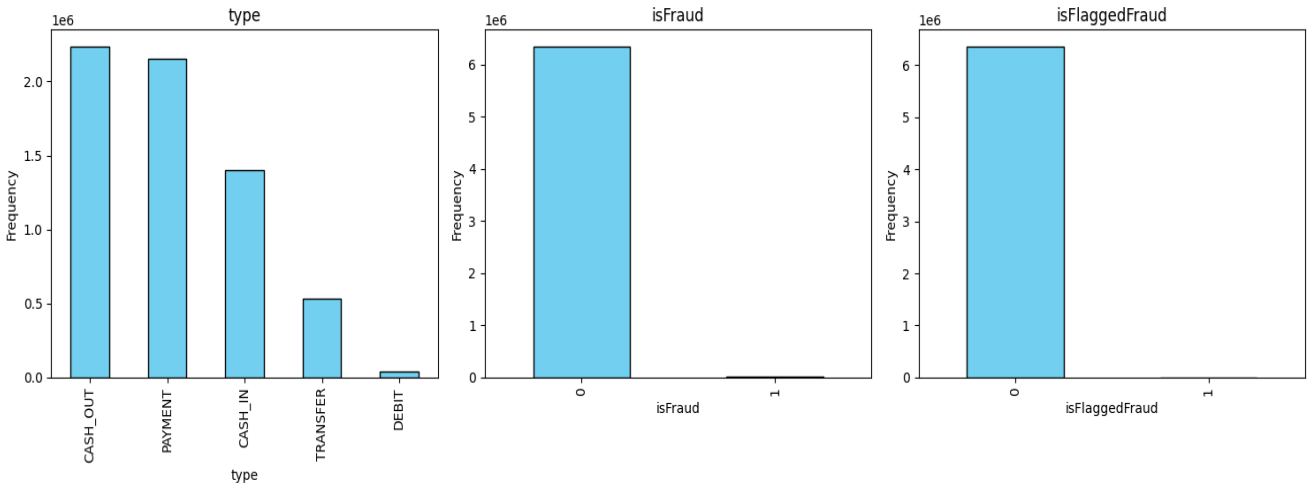
The data visualization process has revealed a significant imbalance in the dataset, with only 0.1% of transactions being identified as fraudulent, while the overwhelming majority are non-fraudulent. To address this imbalance and ensure the robustness of our credit card fraud detection model, I will employ **Synthetic Minority Over-sampling Technique (SMOTE)**. SMOTE is a widely adopted method in machine learning that aims to balance imbalanced datasets by generating synthetic examples of the minority class, in this case, fraudulent transactions. By oversampling the minority class, SMOTE ensures that the model has sufficient representative data to accurately identify fraudulent activities while preventing biases toward the majority class. This approach is essential to enhancing the precision and reliability of our credit card fraud detection system and will be a key component of our research methodology.



The examination of the dataset's numeric variables through histogram plots consistently illustrates a distribution pattern with distinct characteristics. Across these histograms, there is a prominent clustering of data points on the left side, resulting in pronounced peaks, while the right side of the distribution displays notably lower frequencies. This distribution trend suggests that the majority of transactions within the dataset typically involve lower values for these numeric variables, while a minority of transactions exhibit significantly higher values. This observed distribution has implications for the detection of potential outliers and provides valuable insights into transaction behaviors associated with both fraudulent and non-fraudulent activities. Notably, this analysis reveals that a substantial proportion of fraudulent transactions tend to involve smaller transaction amounts, a pattern that might be intentionally employed to avoid raising suspicion during the fraudulent activities.

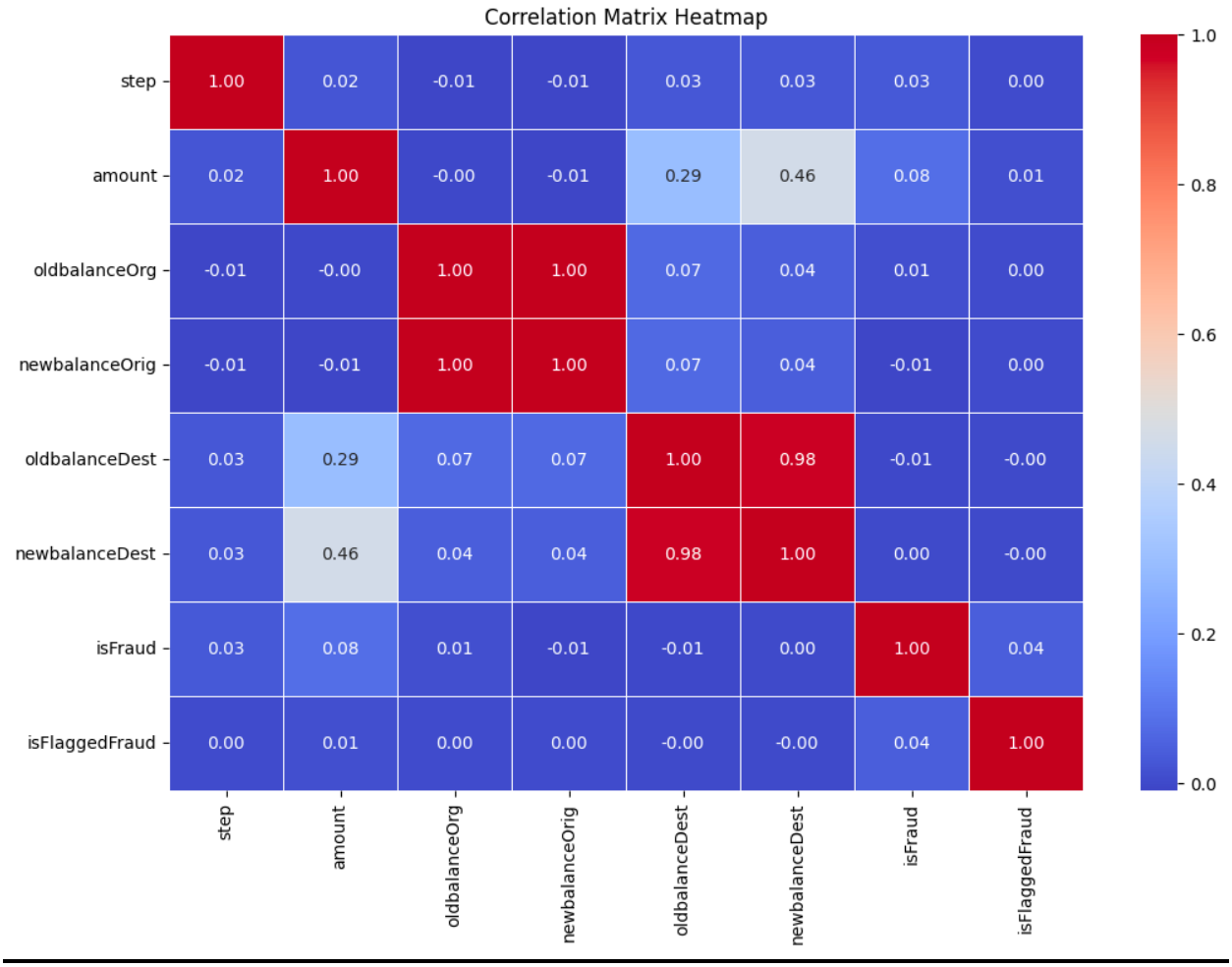


Bar charts were generated to provide a comprehensive visual representation of the categorical variables within the dataset, aiding in a more profound understanding of their characteristics. These bar charts offer insights into the distribution and proportions of different categories within each variable, shedding light on transaction types, fraud labels, and the flagging of potential fraud. This visual exploration not only serves as a means of data comprehension but also highlights any disparities and patterns within these categorical attributes. It is a crucial step in deciphering the role of each categorical variable in the context of credit card fraud detection, contributing to the broader research goal of identifying influential factors and distinct trends that differentiate fraudulent from non-fraudulent transactions.

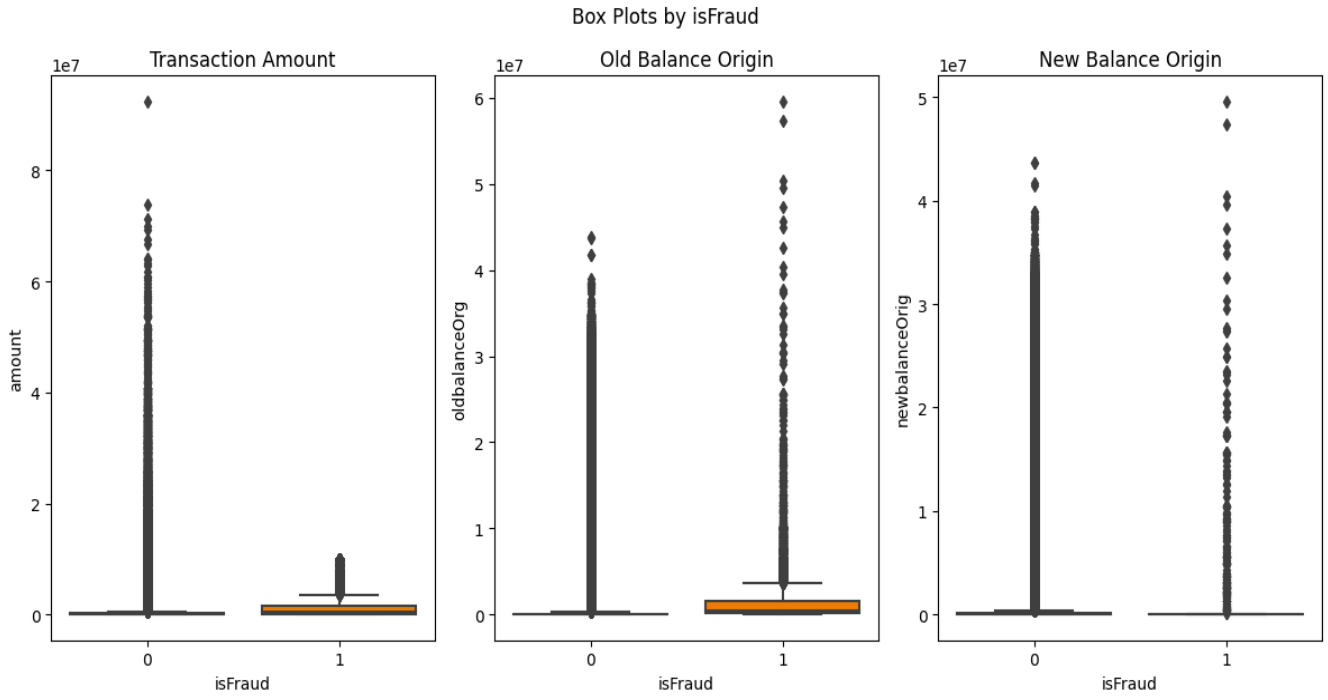


The calculation of correlations among numeric variables has been a pivotal component of this research, enabling a deeper comprehension of the interrelationships between these features. By quantifying these associations, we gain valuable insights into how different numeric attributes influence one another. This endeavor assists in the identification of feature

dependencies and provides a foundation for feature selection, a critical step in developing accurate credit card fraud detection models. It is noteworthy that correlations extend beyond merely determining linear relationships; they uncover the intricate web of connections that can be leveraged to enhance the predictive capabilities of our models. In essence, understanding these correlations is instrumental in our pursuit of crafting models that can effectively differentiate between fraudulent and non-fraudulent transactions, ultimately contributing to the overarching objective of bolstering financial security in the face of credit card fraud.



Box plots have been instrumental in shedding light on the distribution of specific features, namely 'amount,' 'oldbalanceOrg,' and 'newbalanceOrig,' categorized based on the nature of the transactions—fraudulent or non-fraudulent. These visualizations offer a compelling perspective on the spread, central tendencies, and potential outliers within these attributes. The separation by transaction type allows us to discern distinct patterns between fraudulent and non-fraudulent transactions, highlighting potential disparities in these variables based on transaction legitimacy. This nuanced understanding aids in uncovering potential clues that can inform the development of robust credit card fraud detection models. It is worth noting that these box plots offer valuable insights into the distribution characteristics of these features, ultimately contributing to the holistic exploration of transaction behavior and, by extension, enhancing fraud detection techniques. Furthermore, these visualizations reveal that most of the fraud transactions involve smaller amounts, a pattern that can be indicative of an attempt to evade suspicion.



4. TENTATIVE OVERALL METHEDOLOGY

This methodology offers a structured approach from establishing research goals to data collection, analysis, modeling, and ultimately drawing conclusions based on the findings.

Research Goals & Objectives: This initial stage involves defining the research's purpose, outlining the specific goals and objectives, and establishing the overall direction for the project.

Data Collection: This step involves gathering relevant data sources necessary for the research. It encompasses acquiring structured or unstructured data from various sources, ensuring data quality and integrity.

Initial Analysis: This stage involves collecting, cleaning, and comprehending the initial data. It includes defining the problem statement, setting research objectives, and organizing the data for further analysis.

Exploratory Analysis: This phase focuses on examining the data in detail, conducting descriptive statistical analyses, and identifying patterns, trends, and relationships within the data. It aims to gain insights into the characteristics of the dataset before proceeding.

Dimensionality Reduction: This stage aims to reduce the number of variables under consideration. It may involve techniques such as feature selection, feature extraction, or other methods to decrease the complexity of the dataset.

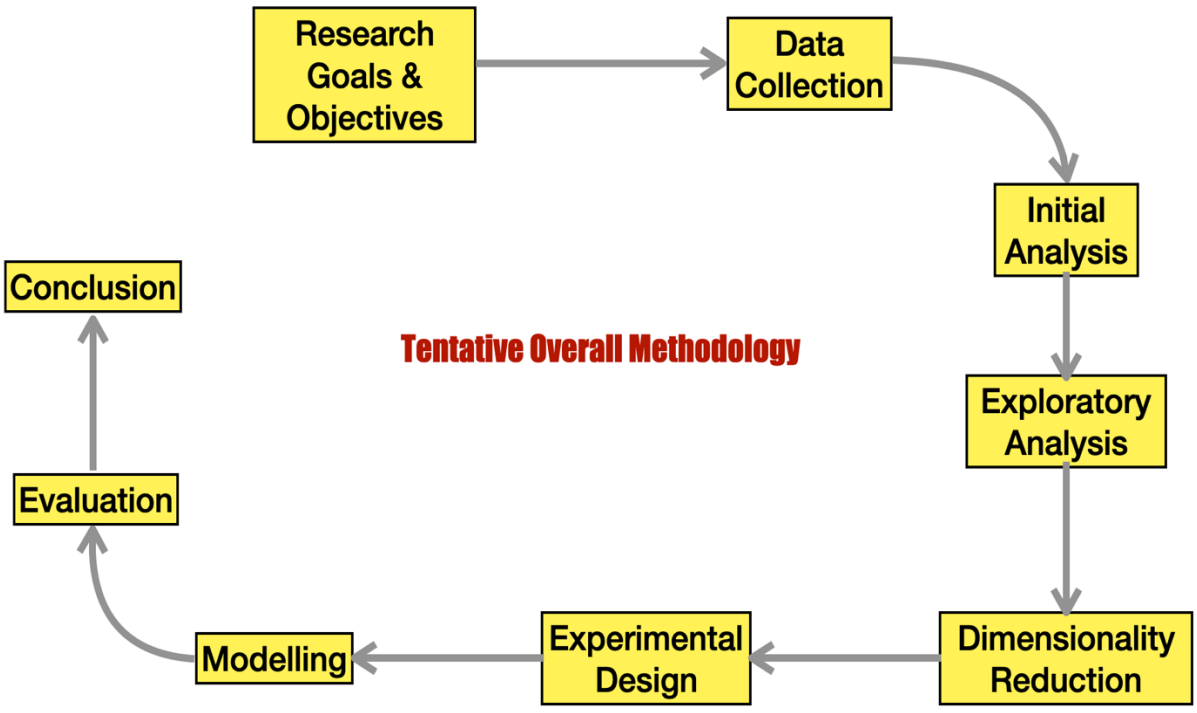
Experimental Design: This phase involves structuring experiments or creating a plan to test hypotheses and validate models. It includes defining the parameters, variables, and methodologies for conducting experiments or simulations.

Modeling: In this stage, core models or algorithms are applied to the data to create a representation of relationships or patterns found during the previous stages. This might involve machine learning, statistical modeling, or other predictive techniques.

Evaluation: After building the models, this phase involves assessing their performance against defined criteria. It includes model validation, testing against various metrics, and comparing different models to select the most effective one.

Improving the Model: Based on the evaluation results, this step focuses on refining or optimizing the models. It involves iterations and enhancements to improve the model's accuracy, robustness, or efficiency.

Conclusions: In this final stage, the research draws conclusions based on the outcomes of the analysis, modeling, and evaluation. This includes summarizing the findings, discussing their implications, and suggesting potential future steps or application



GitHub Repository Link:

https://github.com/gdadu2294/Credit-Card-Fraud-Detection

5. Reference:

[1]

Kaggle Dataset:

https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset

[2]

Jonathan Kwaku Afriyie, Kassim Tawiah, Wilhemina Adoma Pels, Sandra Addai-Henne, Harriet Achiaa Dwamena, Emmanuel Odame Owiredu, Samuel Amening Ayeh, John Eshun.

A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions, *Decision Analytics Journal*, Volume 6, 2023, 100163, ISSN 2772-6622.

https://www.sciencedirect.com/science/article/pii/S2772662223000036

[3]

K.R Seeja and Masoumeh Zareapoor.

FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, *The Scientific World Journal*, vol. 2014, Article ID 252797, 10 pages, 2014.

https://doi.org/10.1155/2014/252797

[4]

Warghade, Swati & Desai, Shubhada & Patil, Vijaykumar. (2020).

Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm. *International Journal of Computer Trends and Technology*. 68. 22-28. 10.14445/22312803/IJCTT-V68I3P105.

https://www.researchgate.net/publication/342391340_Credit_Card_Fraud_Detection_from_Imbalanced_Dataset_Using_Machine_Learning_Algorithm

[5]

Lima, Rafael & Pereira, Adriano. (2017).

Feature Selection Approaches to Fraud Detection in e-Payment Systems. *Lecture Notes in Business Information Processing*. 278. 111-126. 10.1007/978-3-319-53676-7_9.

https://www.researchgate.net/publication/313731885_Feature_Selection_Approaches_to_Fraud_Detection_in_e-Payment_Systems

[6]

Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten.

Feature engineering strategies for credit card fraud detection, *Expert Systems with Applications,* Volume 51, 2016, Pages 134-142, ISSN 0957-4174,

https://www.sciencedirect.com/science/article/pii/S0957417415008386

[7]

Asma Cherif, Arwa Badhib, Heyfa Ammar, Suhair Alshehri, Manal Kalkatawi, Abdessamad Imine.

Credit card fraud detection in the era of disruptive technologies: A systematic review, *Journal of King Saud University - Computer and Information Sciences,* Volume 35, Issue 1,

2023, Pages 145-174, ISSN 1319-1578,

https://www.sciencedirect.com/science/article/pii/S1319157822004062