Credit Card Fraud Detection

CIND820: Capstone Project

Project by: Gaurang Dadu

gdadu@torontomu.ca

Student #501262373

Supervisor: Dr. Ceni Babaoglu

Date of Submission: November 27th, 2023

Ryerson University

Table of Contents

1. ABSTRACT

Credit Card Fraud Detection is a critical challenge in the financial sector, necessitating advanced techniques for precise identification and prevention. This study focuses on employing machine learning models, specifically Random Forest, Decision Tree, and Logistic Regression, within a supervised learning framework to enhance accuracy in detecting credit card fraud. Leveraging a publicly available dataset from Kaggle, the research aims to determine the most effective model in predicting fraudulent transactions.

The primary objective involves training these models using a class variable indicating fraud or non-fraud transactions and evaluating their performance on a separate test dataset, unseen during training. Notably, the dataset is imbalanced, and Synthetic Minority Over-sampling Technique (SMOTE) is applied to address this imbalance. Random Forest emerges as the top-performing model in terms of predictive accuracy.

In the evaluation process, the dataset is split into training and test datasets, and k-fold cross-validation is employed, with random sampling yielding the best results. Methodologies such as Confusion Matrix and F1 Score are applied for robust performance evaluation, offering insights into precision, recall, and overall efficiency in fraud detection. The findings from this research provide valuable guidance to financial institutions in implementing effective fraud prevention measures, contributing to advancements in predictive analytics and data-driven decision-making.

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

2. INTRODUCTION

In response to the increasing prevalence of digital payments and online commerce, this project addresses the critical need to strengthen credit card fraud detection mechanisms. The primary objective is to enhance the precision and efficacy of fraud detection through the deployment of advanced machine learning models.

Credit card fraud, given its complex nature, requires exploration of cutting-edge technologies. This project serves as a comprehensive investigation into the capabilities of three robust machine learning models: Random Forest, Decision Tree, and Logistic Regression. These models, possessing distinct attributes, are strategically chosen for their ability to navigate the intricate landscape of evolving fraudulent activities.

The project leverages a Kaggle dataset containing credit card transactions, serving as the testing ground for the machine learning models. The dataset includes a special class variable distinguishing legitimate transactions from potential fraud, guiding the models through the learning process.

The ultimate challenge for these models lies in their real-world applicability. Beyond theoretical training, the models must accurately predict the legitimacy of new, unseen transactions, ensuring their practical effectiveness in safeguarding financial transactions.

However, a significant obstacle in this endeavor is the imbalanced dataset, where fraudulent transactions are rare compared to legitimate ones. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) is employed. Acting as a reliable aid, SMOTE synthetically creates more instances of the minority class, enabling the models to combat challenges posed by imbalanced data.

To further validate the models' efficacy, strategic techniques such as random sampling and k-fold cross-validation are deployed. These techniques serve as vigilant measures, ensuring that the models not only excel in controlled environments but also exhibit consistent performance when exposed to the unpredictable nuances of real-world data.

This introduction establishes the overarching mission of the project: deploying machine learning models to effectively combat credit card fraud. As the narrative unfolds, we will delve deeper into the technical intricacies, unravel the methodologies employed, and analyze the implications of these models for the future landscape of fraud prevention.

3. LITERATURE REVIEW

In recent years, the surge in credit card fraud has prompted rigorous exploration of sophisticated detection techniques by various researchers. The study conducted by "Jonathan Kwaku Afriyie et al" delves into the performance of logistic regression, random forest, and decision trees for credit card fraud detection. Their comprehensive analysis reveals that random forest emerged as the most effective model, achieving a remarkable accuracy of 96% and an area under the curve (AUC) value of 98.9%. Notably, the research uncovered 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.

In the realm of credit card fraud detection, "K.R. Seeja and Masoumeh Zareapoor" present an innovative approach with "FraudMiner" to address the challenges posed by highly imbalanced and anonymous credit card transaction datasets. The model tackles class imbalance by identifying legal and fraud transaction patterns for each customer through frequent itemset mining, achieving high fraud detection rates and balanced classification rates. This research is a significant contribution, especially in handling anonymous and imbalanced credit card transaction data, offering a valuable perspective for future fraud detection systems.

"Swati Warghade et al" investigated credit card fraud detection challenges, focusing on highly imbalanced datasets. The research explores an array of machine learning techniques and emphasizes the critical role of Synthetic Minority Oversampling Technique (SMOTE) in addressing imbalances. Their work provides a 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, "Rafael Lima and Adriano Pereira" offer a comprehensive exploration into the intricacies of feature selection as a critical aspect of

fraud identification. The 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, culminating in the creation of a robust model for fraud detection.

The research paper authored by "Bahnsen et al" makes a significant contribution to the field of credit card fraud detection by focusing on feature engineering to enhance model accuracy. 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, the research demonstrates the significant impact of these proposed features on fraud detection results, revealing an average increase in savings of 13%.

This comprehensive research paper authored by "Asma Cherif et al" offers a timely and thorough review of credit card fraud detection methodologies from 2015 to 2021. The authors 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, serving as a valuable resource for academic and industrial researchers striving to enhance the efficacy of financial fraud detection systems and develop resilient solutions.

4. RESEARCH QUESTIONS & OBJECTIVE

In the complex landscape of financial transactions, credit card fraud remains a persistent challenge, necessitating advanced detection techniques. This study aims to explore and answer key research questions to contribute valuable insights to the field of credit card fraud detection.

1. Comparative Analysis of Machine Learning Algorithms:

- *Research Question:* How do different machine learning algorithms, specifically Random Forest, Decision Tree, and Logistic Regression, compare in terms of accuracy and efficiency for credit card fraud detection?
- *Objective:* To evaluate and compare the performance of these algorithms in the context of credit card fraud detection, considering accuracy, precision, recall, computational efficiency, and suitability for real-world applications.

2. Impact of Class Imbalance and SMOTE Techniques:

- *Research Question:* What is the impact of class imbalance in credit card fraud datasets, and how can techniques like Synthetic Minority Oversampling Technique (SMOTE) be effectively employed to address this issue?
- *Objective:* To assess the challenges posed by class imbalance in fraud datasets and investigate the effectiveness of SMOTE in mitigating these challenges, aiming for a more balanced and accurate fraud detection model.

3. Contribution to Real-Time Fraud Detection Systems:

- *Research Question:* How can the insights gained from this research contribute to the development of real-time credit card fraud detection systems, and what are the practical implications for financial institutions?
- *Objective:* To identify actionable insights derived from the study that can inform the development and enhancement of real-time credit card fraud detection systems, thereby aiding financial institutions in strengthening their fraud prevention measures.

4. Translating Timed Cross-Validation Results into Recommendations:

- *Research Question:* In what ways can the results obtained from timed cross-validation be translated into actionable recommendations for financial institutions to improve the timeliness and accuracy of their fraud detection systems?
- *Objective:* To interpret the outcomes of timed cross-validation and provide practical recommendations for financial institutions, ensuring the timeliness and effectiveness of their credit card fraud detection systems.

By addressing these research questions, this study seeks to contribute comprehensive insights and practical recommendations to advance the field of credit card fraud detection, offering tangible benefits for financial institutions in safeguarding their systems against fraudulent activities.

5. 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.

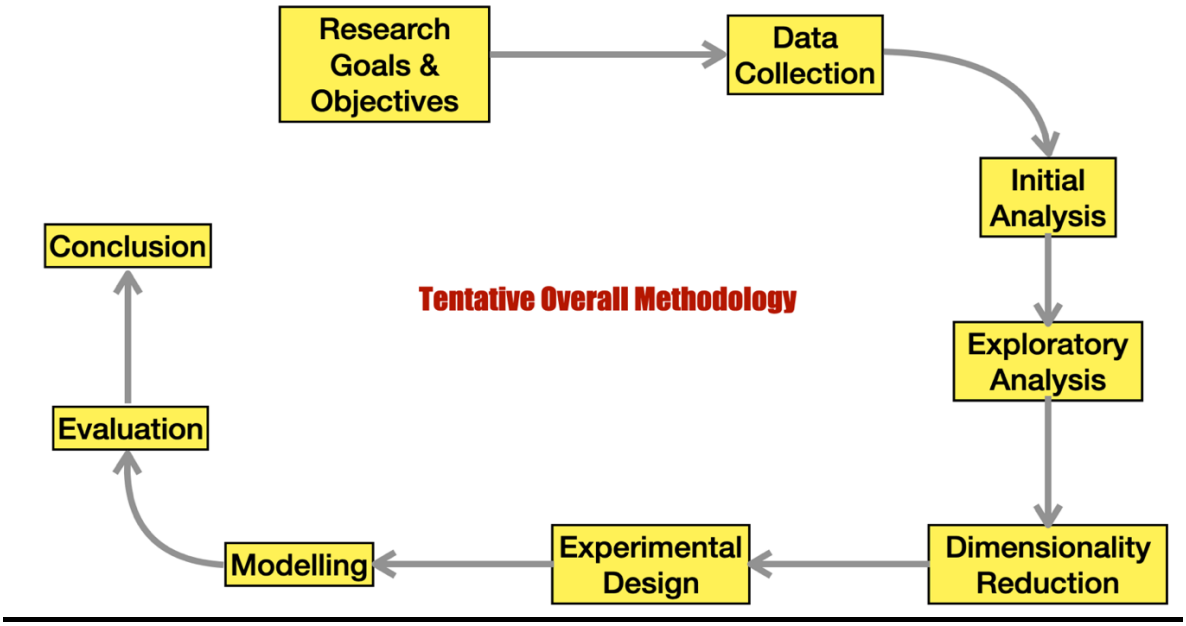
Experimental Design: This phase involves structuring experiments or creating a plan to test hypotheses and validate models. It includes defining the parameters, variables, and methodologies for conducting experiments or simulations.

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



6. DATA DESCRIPTION & INITIAL ANALYSIS

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.

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.

Initial Analysis

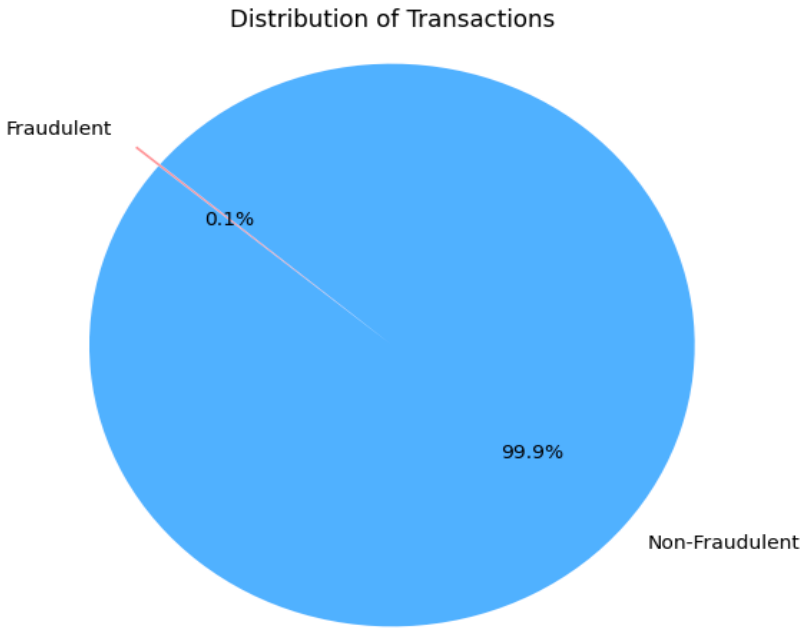
The analysis of the dataset reveals crucial insights into its structure and characteristics. With over 6 million rows and 11 columns, the dataset encompasses various numerical features, including temporal indicators such as 'step,' monetary values like 'amount,' and account balances before and after transactions. Notably, the 'isFraud' column, initially numeric, was transformed into a binary class variable, setting the stage for its role as the target variable in fraud detection classification.

Ensuring data quality is foundational to any analysis, and this dataset excels in that regard. A comprehensive examination uncovered no missing values, and any potential duplicates were

meticulously removed. This meticulous approach to data integrity lays a robust foundation for subsequent analyses and modeling.

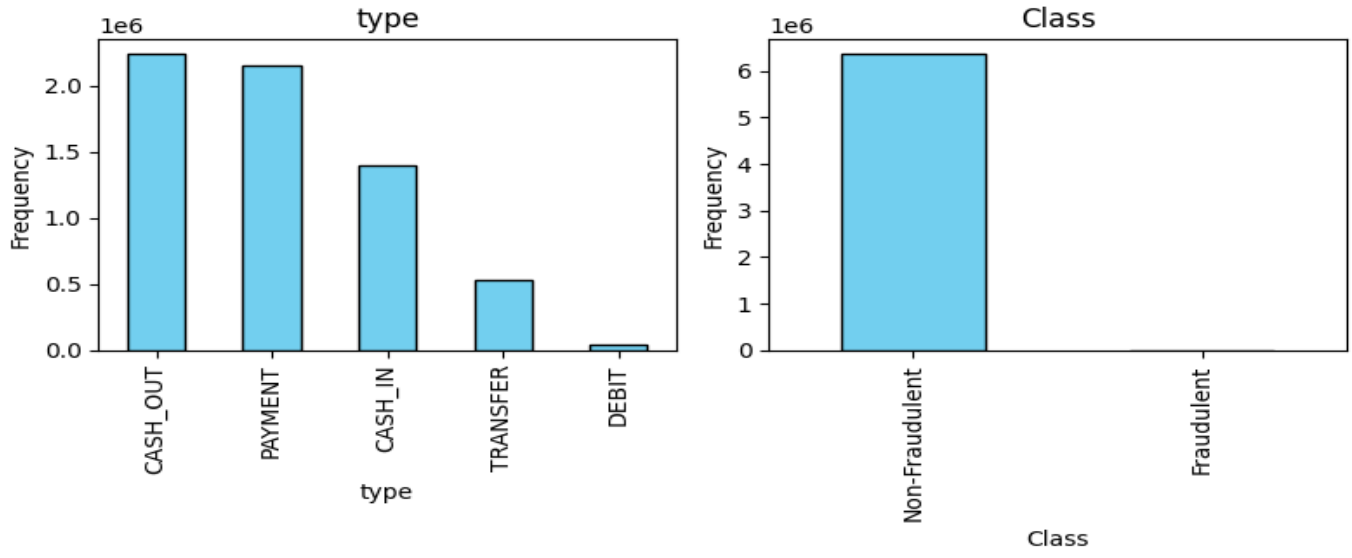
A deeper exploration into the distribution of transaction types provides valuable context. The histogram of the 'type' variable reveals five distinct transaction types: 'CASH_OUT,' 'PAYMENT,' 'CASH_IN,' 'TRANSFER,' and 'DEBIT.' These transactions exhibit varying frequencies, with 'CASH_OUT' dominating the landscape. Understanding the distribution of transaction types is pivotal for grasping the intricacies of financial activities and subsequently informing the fraud detection model.

However - the dataset's pronounced imbalance. A mere 0.1% of transactions are labeled as fraudulent, while the overwhelming majority, 99.9%, are non-fraudulent. This imbalance underscores the need for careful consideration and strategic handling during the subsequent stages of the research. Addressing this imbalance is paramount to ensuring the model's ability to effectively identify and categorize fraudulent transactions.



To address this, Synthetic Minority Over-sampling Technique (SMOTE) will be employed. SMOTE generates synthetic examples of the minority class, ensuring the model has representative data. This strategic oversampling enhances the precision and reliability of our credit card fraud detection system, a pivotal aspect of our research methodology.

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.



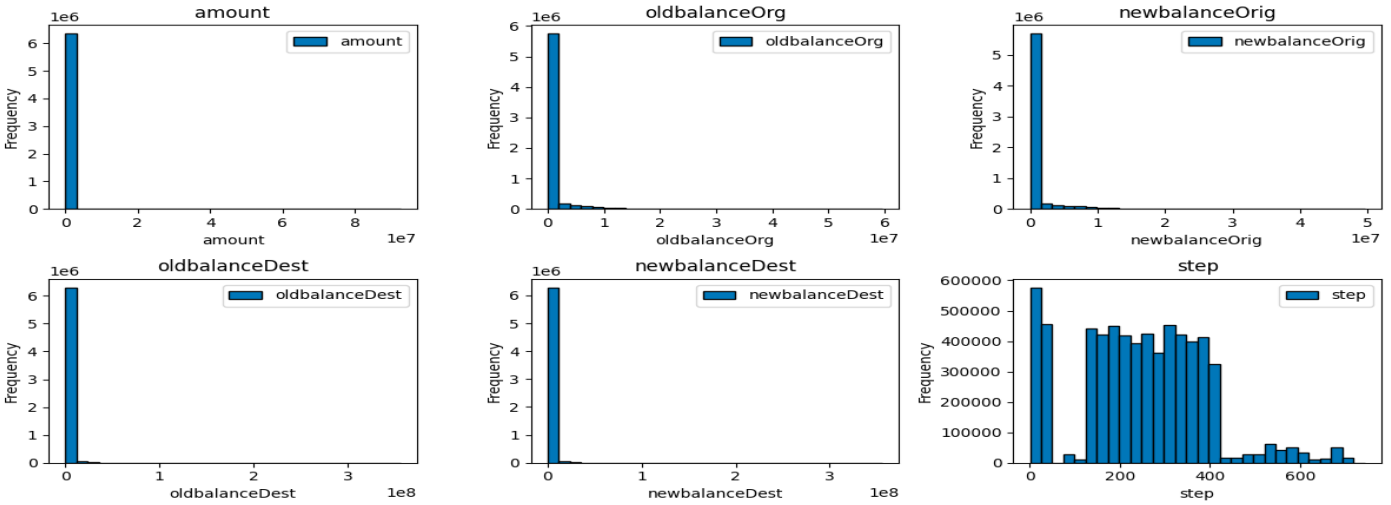
This initial analysis lays a solid foundation for the subsequent phases of the research project. The dataset's characteristics, initial transformations, and the distribution of transaction types provide essential context. The identified imbalance in fraudulent and non-fraudulent

transactions, though challenging, serves as a focal point for the development of strategies and techniques to bolster the effectiveness of the fraud detection model. As the research progresses, these insights will be instrumental in refining methodologies, enhancing model performance, and contributing valuable findings to the domain of credit card fraud detection.

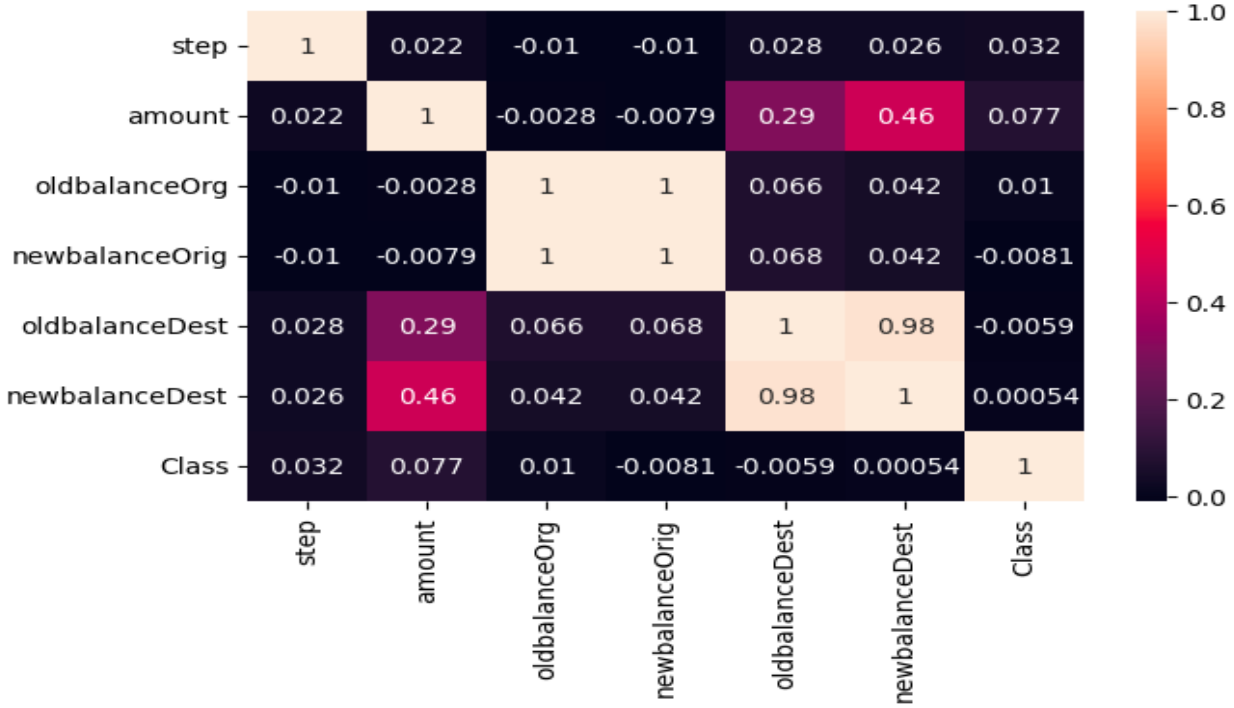
7. EXPLORATORY ANALYSIS

In the exploratory analysis phase, the dataset underwent a refinement process to streamline its features with a keen focus on enhancing credit card fraud detection. The decision-making criterion for this process involved removing seemingly irrelevant columns that did not align with the primary goal of identifying fraudulent activities in credit card transactions. Consequently, 'isFlaggedFraud,' 'nameOrig,' and 'nameDest' were excluded to create a more focused dataset for subsequent analyses.

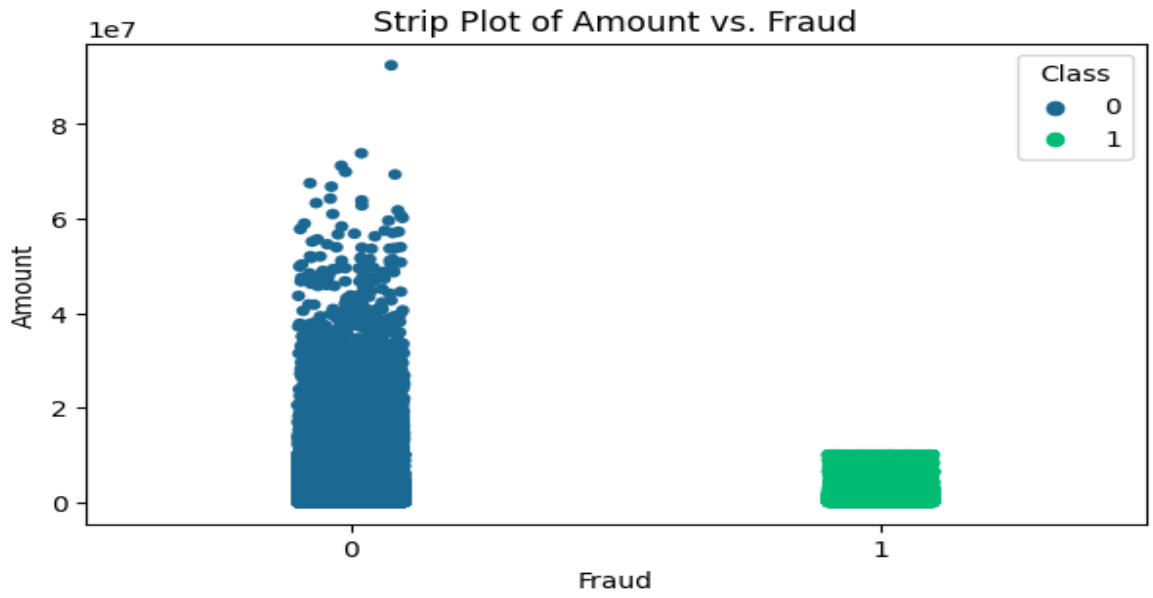
Following this refinement, a thorough examination of the numeric variables was conducted through histogram plots. The insights gained revealed a prevalent leftward skew in the data distributions, indicative of a propensity towards smaller values. Recognizing the impact of data scale on model performance, normalization is scheduled for implementation. This normalization process is particularly crucial for logistic regression, optimizing the data scale for its optimization algorithms. While decision trees and random forest models are generally less sensitive to data scale, normalizing the data could potentially enhance their overall performance. As such, normalization will be incorporated in subsequent stages to ensure an optimized modeling process.



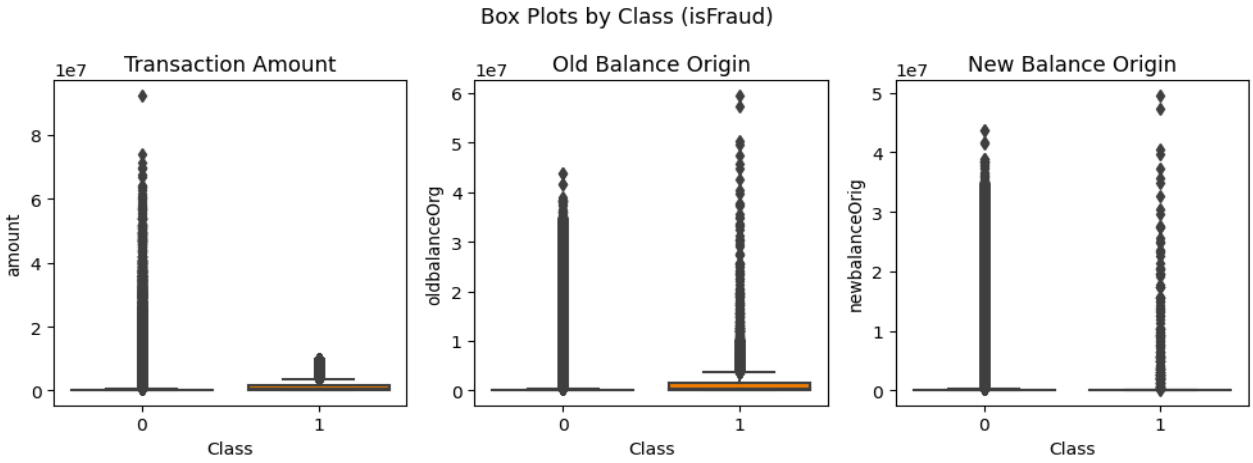
The exploratory analysis also encompassed an investigation into the correlation between variables. Surprisingly, no discernible linear relationships were observed between the 'Class' variable, denoting fraud, and other features. This intriguing finding prompted contemplation on alternative methods for feature selection, as linear correlations seemed insufficient to capture the complexity of the relationships within the dataset.



Further insights emerged from a strip plot visualization focusing on the relationship between transaction amount and fraud exposure. Notably, the visualization revealed a distinct trend – all fraudulent transactions were associated with smaller transaction amounts. This observation implies the potential importance of smaller transaction amounts as a critical feature for identifying fraudulent activities within the dataset, offering valuable insights for the development of effective fraud detection strategies.



To uncover potential outliers and understand the distribution of certain features ('amount,' 'oldbalanceOrg,' and 'newbalanceOrig'), box plots were generated, each segmented by the 'isFraud' variable. The examination highlighted the presence of potential outliers, indicating a potential imbalance in the distribution between fraudulent and non-fraudulent transactions. Caution was exercised in the handling of outliers, considering the scarcity of fraud cases. Direct removal of outliers was deemed inappropriate at this stage, as it could further limit the already sparse number of fraud transactions. Strategies for managing outliers will be carefully considered in subsequent stages, with due consideration for the class imbalance in the dataset.



In conclusion, the exploratory analysis has provided valuable insights into the dataset, shaping the path forward for our credit card fraud detection research. The removal of irrelevant columns has streamlined our focus on key features, enhancing the dataset's relevance to our fraud detection objective. Histograms have illuminated left-skewed data distributions, signaling the need for normalization to optimize model performance, especially for logistic regression.

Notably, the absence of linear relationships between the 'Class' variable and other features has prompted us to explore alternative feature selection methods, acknowledging the complexity of the dataset's dynamics. The strip plot visualization emphasized the potential significance of smaller transaction amounts in identifying fraudulent activities, offering a pivotal clue for refining our fraud detection strategies.

Furthermore, box plots revealed potential outliers and highlighted an unbalanced distribution between fraudulent and non-fraudulent transactions. Careful consideration will be given to handling these outliers in later stages, ensuring an effective balance between outlier removal and preserving the integrity of the limited fraud cases in our dataset.

Moving forward, the next steps include implementing normalization techniques to enhance data scale, exploring alternative feature selection methods beyond linear correlations, and refining outlier handling strategies. These steps are crucial for preparing our dataset for the modeling phase, where machine learning algorithms will be trained to effectively detect credit card fraud. The insights gained from this exploratory analysis will guide the construction of a robust fraud detection model, contributing to the overarching goal of strengthening security measures in financial transactions.

8. DIMENSIONALITY REDUCTION

In the dimensionality reduction stage, critical preprocessing steps were implemented to ready the dataset for further analysis. One-hot encoding of the 'type' variable was performed to ensure compatibility with logistic regression, which requires numerical inputs. While decision trees and random forests can handle categorical data directly, logistic regression specifically necessitates numerical features. The resulting binary values from one-hot encoding aligned with logistic regression's requirements, facilitating effective use across all three models.

Normalization using the robust scaler was deemed essential due to the limited amount of fraud data. Traditional outlier removal methods might lead to the loss of vital information in the already sparse fraudulent transactions. The robust scaler, being less sensitive to outliers, enabled normalization while preserving the integrity of fraud-related data.

Certain columns were converted from float to integer, maintaining the binary classification integrity. Unique values in the 'Class' column were checked, revealing the conversion from

binary integer to float format due to robust scaling. The subsequent conversion back to integer format aimed to avoid confusion, ensuring consistency with the original dataset.

An analysis of the distribution of fraudulent transactions across different transaction types highlighted that fraud is more prevalent in 'type_CASH_OUT' and 'type_TRANSFER' transactions. Notably, 'type_DEBIT', 'type_PAYMENT' and 'type_CASH_IN' categories showed no instances of fraudulent activities.

A no-model evaluation established a baseline performance, resulting in a high accuracy of 99.87% by predicting all transactions as non-fraudulent. This baseline serves as a reference for comparing subsequent model assessments and analyses, providing insights into the inherent distribution of the 'Class' variable within the dataset.

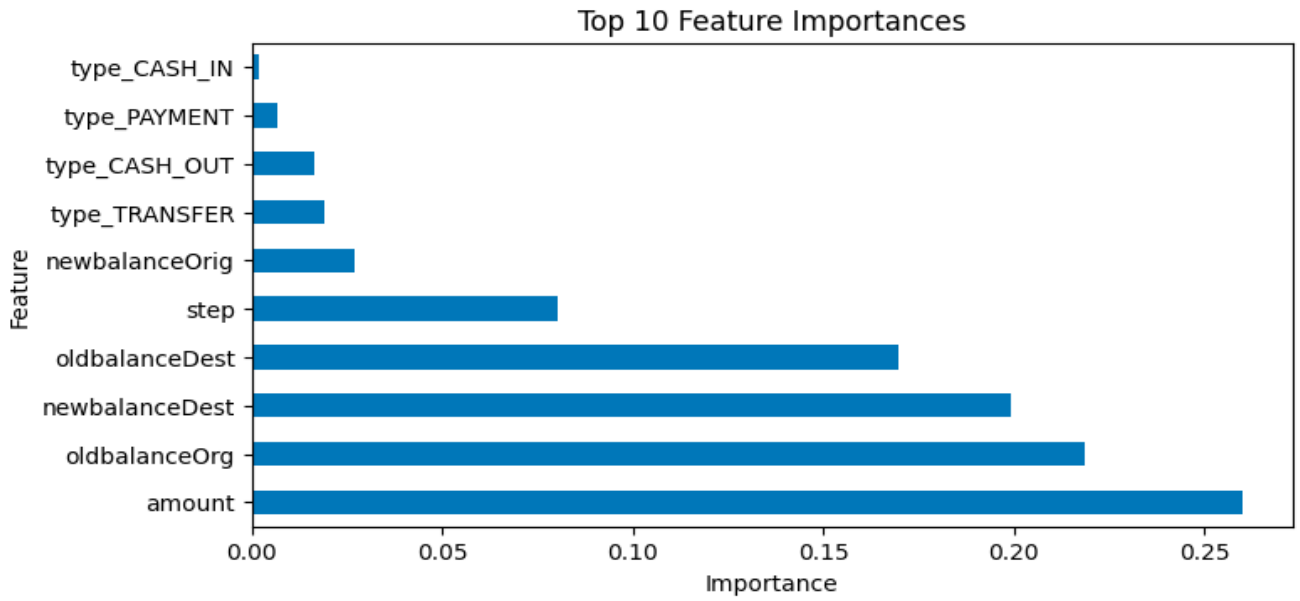
Chi-squared and ANOVA tests were conducted on categorical and numeric variables, respectively. Chi-squared and ANOVA tests are crucial statistical methods employed to assess the relationships between variables.

The Chi-squared test assesses the independence of two categorical variables, such as 'Class' (fraudulent or non-fraudulent) and 'type_' variables (transaction types). The p-value from the test indicates the probability of observing the observed association by chance. A low p-value suggests a significant association between variables, implying that the transaction type is not independent of the occurrence of fraud.

Similarly, the ANOVA test, or analysis of variance, is used to examine the differences in means among multiple groups. In this context, it was applied to numeric variables to evaluate their association with the 'Class' variable. A low p-value from ANOVA indicates a substantial statistical difference in the means of the selected variables, suggesting a strong association with the class distribution (fraudulent vs. non-fraudulent transactions).

The p-value in both tests is a crucial metric. A low p-value (typically below 0.05) signifies that the observed associations are unlikely to occur by chance, indicating the presence of a meaningful relationship between the variables. Conversely, a high p-value suggests that observed associations could be due to random chance, and there may not be a significant relationship between the variables.

Feature importance, determined using a Random Forest Classifier, highlighted the sequence of influential features. Additionally, insights into the occurrence of fraud at different times of the day, reveal the peak hour at 0:00 with 7499 fraud transactions.



While common practice involves removing features with low importance, 'type_CASH_IN' was retained despite its lower importance. This decision ensures that the model incorporates all available information, recognizing the scarcity of relevant features in credit card fraud detection. The detailed preprocessing steps lay a solid foundation for the subsequent modeling phase, aiming to enhance the accuracy and effectiveness of the fraud detection system.

9. EXPERIMENTAL DESIGN

In the experimental design phase of the project, the dataset was split into features (X) and the target variable (y) using random sampling technique. Random sampling involves selecting a subset of data points randomly, without any specific order. In the context of credit card fraud detection, this technique is useful as it helps ensure that the training and test sets represent a diverse range of transactions, mimicking real-world scenarios where fraudulent activities may occur sporadically.

While stratified sampling method was initially considered to maintain the class distribution, it exhibited an unrealistic accuracy of over 99%, which, in a real-world scenario, is unlikely. Consequently, random sampling, which demonstrated higher accuracy during cross-validation and provided more realistic accuracy numbers, was deemed more suitable for the credit card fraud detection project.

The class distribution in the original data, training set, and test set was checked to ensure that the sampling process did not significantly alter the proportion of fraudulent and non-fraudulent transactions. The distributions were found to be consistent across the datasets, maintaining the original imbalance between the two classes.

To address the imbalanced nature of the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training dataset. SMOTE is crucial in the context of credit card fraud detection because it generates synthetic examples of the minority class (fraudulent transactions) to balance the dataset. This technique helps prevent the model from being biased toward the majority class, thereby improving the accuracy of fraud detection. It is essential to note that SMOTE is only applied to the training dataset to ensure the model encounters untouched data during testing, reflecting real-time scenarios.

The 'step' column, initially representing a time variable, was converted into an integer to facilitate the application of SMOTE. The conversion was necessary to maintain the chronological order of transactions in the time series data.

For evaluating the performance of different models—Random Forest, Decision Tree, and Logistic Regression—Time Series Split cross-validation (TS-CV) was employed. TS-CV proved particularly effective for time-dependent data, preserving the temporal order during cross-validation. While k-fold cross-validation was also considered, TS-CV exhibited better results, making it the preferred choice in this case.

This approach is essential in fraud detection, especially considering that most fraud occurrences were concentrated at 0:00 hours. By leveraging TS-CV, the model can focus on specific times, enhancing its ability to detect fraudulent transactions during these critical periods.

The cross-validation results indicated varying levels of accuracy for each model. The Random Forest model achieved an average accuracy of approximately 96.6%, the Decision Tree model averaged around 97.99%, and the Logistic Regression model exhibited an average accuracy of 85.35%. These scores reflect the models' ability to correctly classify transactions as fraudulent or non-fraudulent during the cross-validation process.

It is important to note that, in this context, high accuracy alone does not necessarily signify the superiority of a model. Additional metrics such as precision, recall, and F1 score are crucial for a comprehensive evaluation. Precision measures the accuracy of positive predictions, recall gauges the ability to capture all positive instances, and the F1 score balances precision and recall. Considering these metrics alongside accuracy will provide a more nuanced understanding of the model's effectiveness.

The logistic regression model underwent standard scaling because it was experiencing convergence issues during optimization. This occurred because the ***lbfgs**** optimization algorithm was struggling to find the best solution within the given number of iterations. The problem arose due to variations in the scales of the input features. Standard scaling was introduced to bring all features to a common scale, making them comparable and preventing convergence problems. In simpler terms, it's like ensuring everyone speaks the same language—standard scaling helps the logistic regression model understand and optimize the features more effectively, improving its overall performance.

In conclusion, the strategic choice of random sampling, the incorporation of SMOTE to address imbalanced data, and the application of TS-CV lay a strong foundation for developing an effective credit card fraud detection system. While accuracy provides valuable insights, a holistic evaluation using multiple metrics will guide the selection of the most suitable model for further refinement and evaluation in subsequent phases of the project.

Lbfgs - *The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) is an optimization algorithm used for solving unconstrained optimization problems.*

In the context of logistic regression, the algorithm works to find the optimal parameters that minimize the logistic loss function. It iteratively adjusts the parameters based on the gradient of the loss function, attempting to reach the minimum of the function where the model is most accurate. The warning encountered suggests that the LBFGS algorithm reached the maximum number of iterations without converging to a solution, and scaling the data helps mitigate this issue.

10. MODELLING

In the modeling phase of the project, three distinct machine learning models were employed to tackle the credit card fraud detection challenge. The models selected were Random Forest, Decision Tree, and Logistic Regression, each chosen for its unique strengths and characteristics.

To initiate the modeling process, the preprocessed training data was utilized. This data had undergone essential preprocessing steps, including random sampling for splitting, Synthetic Minority Over-sampling Technique (SMOTE) application to address imbalanced data, and necessary conversions such as transforming the 'step' column into an integer to maintain the chronological order.

Random Forest, a powerful ensemble learning method, was employed as the first model. This model operates by constructing a multitude of decision trees during training and outputs the mode of the classes for classification problems. Cross-validation was conducted using Time Series Split, allowing the model to be evaluated in a time-dependent manner, preserving the temporal order of transactions.

The Decision Tree model, a simpler yet effective algorithm, was the second model in the lineup. Decision trees partition the data into subsets based on features, making them adept at capturing complex relationships. Similar to Random Forest, this model underwent Time Series Split cross-validation to assess its performance over time.

Lastly, Logistic Regression, a fundamental algorithm for binary classification, was included as the third model. Logistic Regression models the probability that an instance belongs to a particular category, making it suitable for predicting the likelihood of fraud in credit card

transactions. Standard scaling was applied to the features to address convergence warnings from the LBFGS optimization algorithm.

Cross-validation scores were obtained for each model, providing insights into their average accuracy over different time intervals. The Random Forest model exhibited an average accuracy of approximately 96.6%, the Decision Tree model achieved around 97.99%, and the Logistic Regression model showed an average accuracy of 85.35%. While accuracy is a crucial metric, further evaluation using precision, recall, and F1 score will be essential to determine the overall effectiveness of each model.

This modeling phase represents a critical step in the development of the credit card fraud detection system. The diverse set of models enables a comprehensive understanding of their strengths and weaknesses in addressing the specific challenges posed by imbalanced and time-dependent transaction data. The results obtained will guide further refinement and optimization in the subsequent stages of the project.

11. EVALUATION

In the evaluation phase, the models—Random Forest, Decision Tree, and Logistic Regression—were assessed using the test dataset, providing a comprehensive understanding of their performance.

The Random Forest model demonstrated exceptional accuracy, achieving 99.91%. Precision, representing the proportion of true positive predictions among all positive predictions, stood at 60.27%. The recall, indicating the proportion of actual positive instances correctly predicted, was notably high at 95.12%. The F1 score, a harmonic mean of precision and recall, reached 73.79%, reflecting a balanced performance. The confusion matrix revealed that out of over 1.2

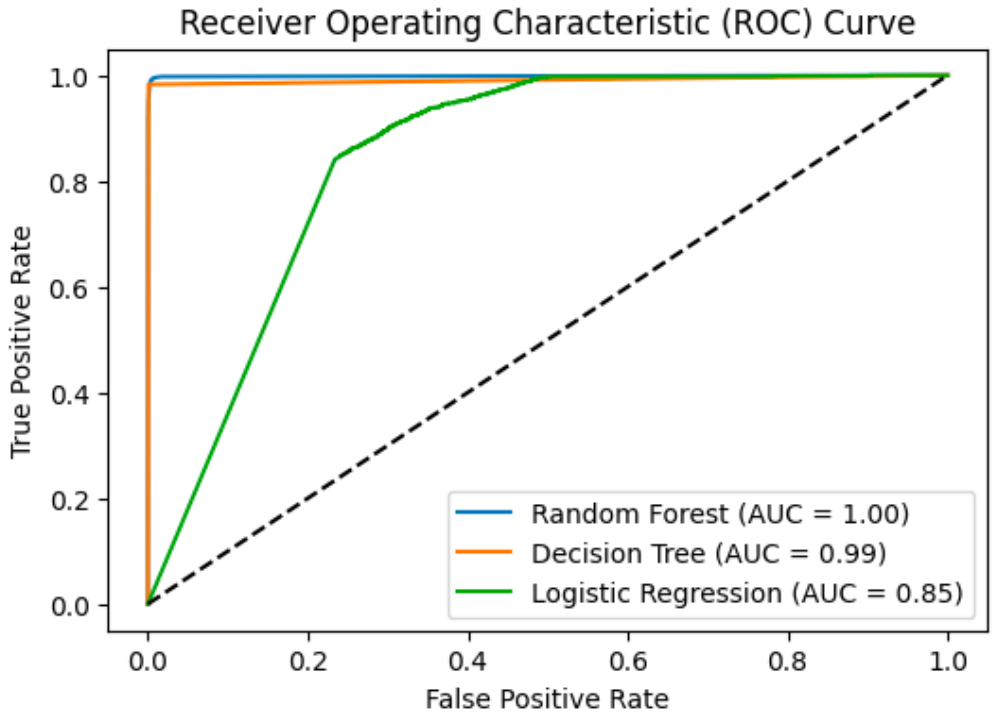
million non-fraudulent instances, only 1,016 were misclassified, while 79 out of 1,620 fraudulent instances were not predicted. The Area Under the Curve (AUC) in the Receiver Operating Characteristic (ROC) curve analysis was an impressive 99.86%, confirming the model's robust ability to discriminate between fraudulent and non-fraudulent transactions.

Similarly, the Decision Tree model exhibited a high accuracy of 99.93%, with precision, recall, and F1 score reaching 66.31%, 98.27%, and 79.18%, respectively. The confusion matrix indicated superior performance, misclassifying only 809 non-fraudulent instances out of over 1.2 million and 28 fraudulent instances out of 1,620. The AUC for the Decision Tree model was commendable at 99.10%, affirming its strong discriminatory capability.

In contrast, the Logistic Regression model displayed lower accuracy at 65.59%, primarily due to a precision of 0.34%, which signifies a large number of false positives. The recall, however, was 93.09%, suggesting that the model effectively identified a substantial proportion of fraudulent instances. The F1 score, being the harmonic mean, was relatively low at 0.68%. The confusion matrix indicated that the model struggled with a high number of false positives (437,798) compared to true positives (1,508). The AUC for Logistic Regression was 84.58%, reflecting a moderate ability to distinguish between the two classes.

The analysis of ROC curves further emphasized the model's discriminatory power. The Random Forest and Decision Tree models exhibited AUC values of 99.86% and 99.10%, respectively, signifying their excellent performance. The Logistic Regression model, with an AUC of 84.58%, demonstrated comparatively lower discriminative ability.

In summary, the evaluation phase provided valuable insights into the strengths and weaknesses of each model. While Random Forest and Decision Tree models displayed high accuracy and AUC, Logistic Regression showed limitations, especially in terms of precision. These findings will guide further refinement and optimization efforts in subsequent stages of the project.



12. LIMITATIONS

The research project encountered several limitations that influenced the scope and depth of the analysis. One significant limitation pertained to the dataset's attributes; although the dataset encompassed a vast number of transactions, the available features were relatively limited. This constrained the ability to engage in comprehensive feature engineering, hindering the extraction of more nuanced insights and patterns necessary for building a highly robust model. The scarcity of relevant features underscored the challenge of developing a more sophisticated understanding of the factors contributing to credit card fraud.

Another notable limitation revolved around the highly imbalanced nature of the dataset. Addressing class imbalance is a critical aspect of fraud detection models, and while Synthetic Minority Over-sampling Technique (SMOTE) was employed to alleviate this issue, there exist

additional techniques that could be explored in future research. Methods such as under-sampling the majority class, using different sampling algorithms, or employing advanced ensemble techniques specifically designed for imbalanced datasets could potentially enhance the model's ability to generalize across both classes.

Limited computing power posed a practical constraint on the project. With a large dataset and computationally intensive models, such as Random Forest, the computational resources required became a crucial consideration. The trade-off between model complexity and computational efficiency needed careful consideration to ensure feasible execution times. This highlights the importance of optimizing code, utilizing parallel computing where possible, and exploring alternative algorithms that strike a balance between accuracy and computational demands.

The temporal aspect of the 'step' feature added a layer of complexity to the analysis. While the 'step' represented time in hours, capturing only daily information limited the depth of temporal insights. Future data collection endeavors should prioritize gathering additional temporal details, such as day, month, and year, to facilitate a more comprehensive understanding of patterns over time. Expanding the temporal dimension would likely enhance the model's ability to discern temporal patterns in credit card transactions, potentially improving fraud detection accuracy.

Furthermore, the overarching need for more relevant information in the dataset emerged as a general limitation. As fraud detection inherently relies on a multifaceted understanding of transactional patterns, incorporating additional relevant features and contextual information is crucial. Future iterations of the research should explore ways to augment the dataset with pertinent variables that offer deeper insights into the dynamics of credit card transactions, ultimately contributing to the development of more effective fraud detection models.

13. CONCLUSION

In conclusion, this research project endeavors to unravel the complexities of credit card fraud detection, striving to answer pivotal questions and provide practical insights for advancing fraud prevention efforts. A detailed comparative analysis of machine learning algorithms—Random Forest, Decision Tree, and Logistic Regression—unveiled intriguing patterns. Surprisingly, the Decision Tree model emerged as the top performer, showcasing superior accuracy and efficiency in the intricate landscape of credit card fraud detection.

The investigation into the impact of class imbalance on fraud datasets highlighted the critical role of dataset balance. Notably, the Synthetic Minority Oversampling Technique (SMOTE) surfaced as a crucial strategy, effectively countering class imbalance and fortifying the accuracy of fraud detection models. These findings underscore the necessity of addressing imbalances to ensure the robustness and fairness of machine learning models in fraud prevention. A crucial test was conducted, revealing that cross-validation without applying SMOTE led to unrealistic high accuracy, indicating bias towards one class.

The focus on real-time fraud detection systems offers actionable implications for financial institutions. Practical insights derived from the research provide tangible recommendations for strengthening real-time credit card fraud detection systems. Financial institutions can leverage these insights to fortify their defenses, respond promptly to potentially fraudulent activities, and enhance overall system efficiency.

The introduction of timed cross-validation added a temporal dimension to model evaluation, emphasizing the significance of considering time-dependent patterns in fraud detection. The Decision Tree model's consistent performance across various metrics, including accuracy and

efficiency, positions it as an optimal choice for financial institutions seeking to augment their credit card fraud detection systems.

In summary, this research project not only contributes to the theoretical understanding of credit card fraud detection challenges but also provides practical recommendations for implementing effective fraud prevention strategies. The Decision Tree model's superior performance underscores its potential as a preferred algorithm for financial institutions aiming to bolster the security and accuracy of credit card transactions. As the financial landscape evolves, the insights gleaned from this study serve as a valuable guide for future research and the ongoing refinement of fraud detection methodologies.

While this research project has provided valuable insights into credit card fraud detection, it is important to acknowledge its inherent limitations and opportunities for improvement. The dataset's limited attributes posed a challenge, restricting the depth of feature engineering and insights extraction. Overcoming this limitation would involve acquiring datasets with richer attributes, enabling more comprehensive analyses and the identification of nuanced patterns.

The highly imbalanced nature of the dataset emphasized the need for advanced techniques to address class imbalance. While SMOTE proved effective, exploring additional methods, such as Adaptive Synthetic Sampling (ADASYN) or borderline-SMOTE, could offer further enhancements. Additionally, obtaining more relevant features specific to credit card transactions would contribute to a more holistic understanding of fraud patterns.

The constraint of limited computing power surfaced as another limitation, impacting the efficiency of certain algorithms, particularly on large datasets. Future iterations of this study could benefit from advanced computing resources, potentially exploring distributed computing or cloud-based solutions to overcome these constraints and enable the exploration of more complex models.

Moreover, the reliance on the 'step' feature, representing time in hours for a single day, underscores the importance of collecting more comprehensive temporal information in future datasets. Including details such as day, month, and year would facilitate a more nuanced analysis of temporal patterns in fraud occurrences.

As with any research endeavor, this study provides a foundation for future exploration and refinement. Building upon this foundation requires the integration of more in-depth knowledge, incorporating advancements in machine learning, and leveraging practical experiences in fraud detection.

By continually addressing these limitations and embracing evolving methodologies, future research in credit card fraud detection can make substantial strides toward developing even more robust and effective models.

14. 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

GitHub Repository Link:

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