Credit Card Fraud Detection

CIND 820 Capstone Project

By: Anjali Verma

Student Number :501223571.

Supervisor: Tamer Abdou

Submission Date: November 30,2023

**Table of Contents\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**1. Abstract ……………………………………………………………………………..... 3**

**2. Introduction ………………………………………………………………………........ 4**

**3. Literature Review……………………………………………………………………… 6**

**4. Research Questions and Objective…………………………………………………...... 8**

**5. Overall Methodology.………………………………………………..................……… 9**

**6. Data Description and Initial Analysis ......................................................................... 11**

**7. Exploratory Analysis…………………………………………………………………...15**

**8. Dimensionality reduction………………………………………………………………18**

**9. Splitting………………………………………………………………………..................19**

**10. Modelling……………………………………………………………………………….21**

**11. Evaluation………………………………………………………………………………22..**

**12. Limitations………………………………………………………………………………25**

**13. Conclusion………………………………………………………………………………26**

**14. Reference .......................................................................................................................27**

**Abstract\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

With the rapid growth of credit card usage, credit card fraud has become a serious issue costing billions of dollars each year. Advanced fraud detection systems will be needed to identify fraudulent transactions. This research aims to develop an automated credit card fraud detection system using machine learning techniques.Multiple supervised machine learning algorithms will be implemented, including logistic regression, Decision Tree, random forest and Naïve Bayes. The models will be trained on the dataset to classify transactions as either fraudulent or valid. The best performing model will be integrated into a fraud detection application.

The primary goal 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 time series k-fold cross-validation is employed, with random sampling yielding the best results.Performance measures such as Confusion Matrix and F1 Score are applied for Providing perspectives 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.

**Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Credit card fraud detection is an essential component of modern financial security systems that employ state-of-the-art technology to shield individuals and businesses from fraudulent activity. Strong fraud detection systems are necessary because cyber threats are growing along with this study's dependence on electronic transactions. This field encompasses a wide range of techniques, ranging from traditional rule-based systems to cutting-edge artificial intelligence and machine learning algorithms. Recent changes in the credit card fraud landscape necessitate the use of flexible and adaptive detection techniques. Since the introduction of EMV chip technology, fraudsters are concentrating more on online and digital channels because it is now harder to clone physical cards.

Behavioral analytics, anomaly detection, and real-time transaction pattern analysis are some of the adaptations that fraud detection systems have had to make. The integration of machine learning and data analytics is essential for spotting anomalies and patterns that point to fraudulent activity. Predictive models are utilized by institutions to stay ahead of evolving tactics by continuously learning and adapting to emerging threats. This nexus of technology and finance highlights the need for constant vigilance in a dynamic environment where fraud detection innovation is critical to preserving the integrity of financial transactions and gaining the confidence of both consumers and companies.

The project makes use of a Kaggle dataset that encompasses credit card transactions, serving as the experimental arena for testing machine learning models. Within this dataset, a unique class variable is present, differentiating legitimate transactions from potential fraud. This variable guides the models through their learning process, enhancing their ability to discern and classify transactions accurately.  
A notable challenge in this undertaking is the imbalanced dataset, wherein instances of fraudulent transactions are scarce compared to legitimate ones. To tackle this issue, we employ the Synthetic Minority Over-sampling Technique (SMOTE).

To enhance the validation of the models' effectiveness, strategic methods such as random sampling and k-fold cross-validation are implemented.

This introduction sets the central objective of the project: employing advanced machine learning models to proactively counteract credit card fraud and enhance transaction security.

**Literature Review**

In this project, I collected and analyzed a number of research papers published. Authors of these research papers applied various approaches to their research problems. They have applied different machine learning algorithms and compared the accuracy of the models to identify the most effective one.

**Related studies**

In the study of "Jonathan Kwaku Afriyie et al" ,the performance of three different machine learning models: logistic regression, random forest, and decision trees to classify, predict, and detect fraudulent credit card transactions. They compared these models’ performance and show that random forest produces a maximum accuracy of 96% (with an area under the curve value of 98.9%) in predicting and detecting fraudulent credit card transactions.

Another comparative study by G. Niveditha, K. Abarna, and G. v. Akshaya et al investigates different classification algorithms for highly skewed dataset namely logistic regression, random forest, decision trees, and naïve Bayes. According to the results, the random forest classifier has the best performance with an accuracy of 96.77%, precision of 100%, recall of 91.11%, and F1 score of 95.43%.

According to Jain et al. (2020), the two transfers where frauds are most prevalent are Cash Out and the ones where money is transferred to a merchant before being transferred to users or occasionally, unknowingly, to fraudsters. The first transfer involves money being transferred from one user to another, a fraudster, or a customer. The second transfer is where frauds are most prevalent. In their assessment of various machine learning algorithms for the identification of frauds when using credit cards.

In another study by “Ozlem” to handle the imbalanced data, SMOTE (The Synthetic Minority Oversampling Technique) technique was used to overcome this problem.

In the realm of electronic transaction fraud detection, the extensive research conducted by "Rafael Lima and Adriano Pereira" delves into a thorough exploration of feature selection as a crucial component of fraud identification. The study rigorously assesses the influence of class imbalance on feature selection, shedding light on the inadequacies of traditional methods in effectively identifying anomalies, particularly within highly imbalanced datasets.

John, Adebayo and Samuel et al have used Nave Bayes and K-Nearest Neighbourhood and logistic regression algorithms were developed, and the implementation of these algorithms has been done in Python. In order to solve the problem of data unbalancing, they have used oversampling and undersampling techniques, so the imbalance dataset will be converted into two datasets. The algorithms performances have been evaluated on the basis of various metrics.

Hence after reading the papers, I observe that solving this problem has many approaches, and I will also apply three different machine learning models: logistic regression, random forest, and decision trees to classify, predict, and detect fraudulent credit card transactions. and every model applied would lead to prediction with different accuracies from those evaluation metrics I’ll predict the best classifier to detect fraudulent transactions.

**Research Questions**

* Which machine learning algorithms (logistic regression, SVM, random forest, etc.) are most effective at detecting different types of fraudulent credit card transactions?
* What are the most important transaction features that allow machine learning models to accurately distinguish between fraudulent and valid credit card transactions?
* Which transaction type exhibits higher incidence of fraudulent transactions ?

## Objectives

* To analyze how well random forest, Decision Tree, logistic regression and Naïve Bayes detect credit card fraud.
* To examine transaction characteristics for their effect on model accuracy, including transaction amount, location, and time.
* To examine which kind of transaction—online or in-store, for example—shows the highest frequency of fraudulent transactions.

To provide refined machine learning techniques to improve the identification of credit card fraud.

**Github link :**

<https://github.com/A100Verma/CIND-820-Capstone-Project>

**Overall Methodology**

This methodology presents a systematic framework, guiding the research process from establishing initial goals to concluding with data collection, analysis, modeling, and deriving conclusions based on the findings.

1.**Clarification of Research Goals & Objectives:**

In this preliminary phase, the research's purpose is distinctly outlined, objectives are detailed.

2.**Acquisition of Data:**

This step involves compiling relevant data from diverse sources essential for the research. 3.**Initial Data Analysis:**

This stage entails the collection, cleansing, and comprehension of the initial dataset. 4. **Exploration of Data Patterns:**

During this phase, a meticulous examination of the data occurs. Descriptive statistical analyses are conducted to unveil patterns, trends, and relationships within the dataset. The objective is to gain insights into the dataset's characteristics before progressing.

5. **Dimensionality Reduction :**

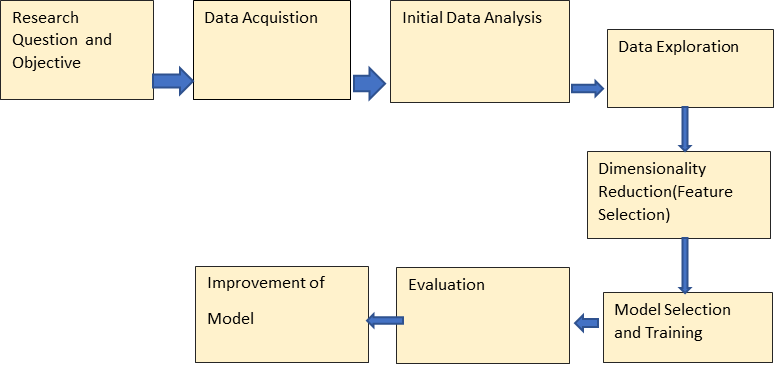
This stage concentrates on minimizing the number of variables considered. Techniques like feature selection, feature extraction, or other methods are applied to simplify the dataset.

6. **Model Selection and Training :**

In this stage of a research methodology, the focus is on applying core models or algorithms to the data with the objective of creating a representation of relationships or patterns identified in the earlier stages of the research process.

7. **Evaluation:.** It involves validating the models, testing them using various metrics, and comparing different models to choose the most effective one.

8. **Improving the Model:**

Building on the evaluation results, this step aims to enhance and optimize the models. It involves making iterations and enhancements to boost the model's accuracy, robustness, or efficiency.

**Descriptive Statistics**

The dataset I'm working with is called the "Online Payments Fraud Detection Dataset."

This dataset is found on Kaggle link

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

It contains a substantial 6,362,620 rows of data, with each row representing a different online transaction.

The dataset has 11 features - step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, and isFraud and isFlagged Fraud.

'step' represents the time of transaction in hourly intervals, means step one equals to one hour.

'type' denotes the transaction type - CASH\_OUT, PAYMENT or TRANSFER.

'amount' indicates the transaction amount.

‘nameOrig' and 'nameDest' contain anonymized alphanumeric identifiers for the sender and recipient respectively.

The old and new balance variables capture the account balances before and after the transaction.

The 'isFraud' variable is the target with 1 indicating a fraud transaction and 0 denoting a legitimate transaction. This reveals a highly imbalanced class distribution with just 0.172% fraudulent transactions.

‘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.

Table

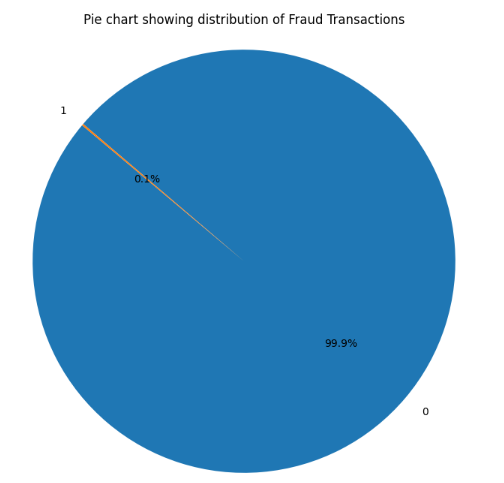
|  |  |  |
| --- | --- | --- |
| Features | Description | Type |
| Step | Maps a unit of time in the real world. In this case 1 step is 1 hour of time. | Numerical |
| type | CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER | Categorical |
| amount | Amount of the transaction in local currency. | Numerical |
| nameOrig | customer who started the transaction | Nominal |
| oldbalanceOrg | initial balance before the transaction | Numerical |
| newbalanceOrig | new balance after the transaction | Numerical |
| nameDest | customer who is the recipient of the transaction | Nominal |
| oldbalanceDest | Initial balance recipient before the transaction. | Numerical |
| newbalanceDest | New balance recipient after the transaction | Numerical |
| isFraud | This is the transactions made by the fraudulent agents,is fraud or not fraud | Categorical  (Nominal) |
| isFlaggedFraud | The business model aims to control massive transfers from one account to another and flags illegal attempts. | Categorical  (Nominal) |

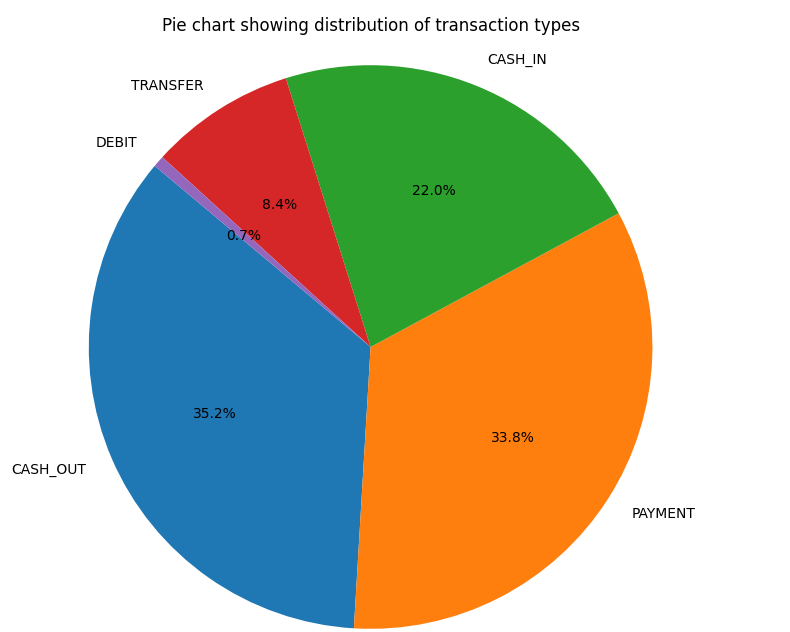
The online payments fraud detection dataset contains 6,362,619 transactions with 11 variables. The target variable 'isFraud' indicates whether a transaction is fraudulent or not, with 1 denoting fraud and 0 denoting a valid transaction.

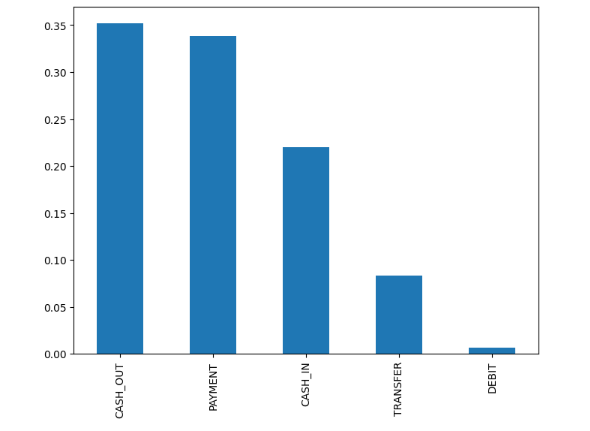
It shows that 99.9% are non-fraudulent transactions and only 0.01% are fraudulent. so, the target variable is highly imbalanced. The dataset exhibits class imbalance with the fraud cases (isFraud = 1) accounting for just 0.01% of the transactions. This skew is common in fraud detection datasets since illegitimate activities occur relatively infrequently compared to regular day-to-day transactions.

To tackle this issue, we'll use SMOTE (Synthetic Minority Over-sampling Technique). SMOTE creates synthetic instances of the minority class, ensuring our model has diverse and representative data. This strategic oversampling improves the accuracy and dependability of our credit card fraud detection system—a crucial element of our research approach.

**Initial Analysis**

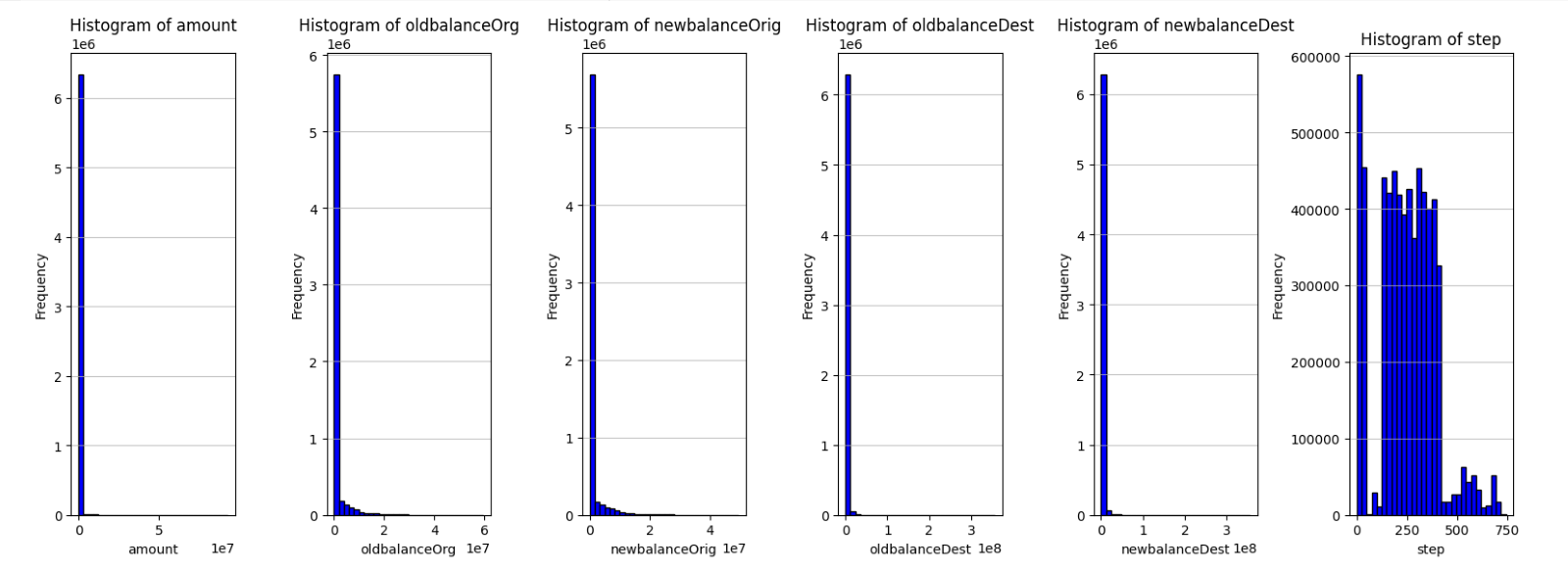


Looking at the transaction types, 35% are CASH\_OUT, 34% are PAYMENT while 31% are other transaction types.

  
The figure shows how much cash removed from a ledger in the US, isolated into various classifications, including cash, credit, debit, and transfers. The biggest classification is cash out and transfer types.

**Exploratory Analysis**

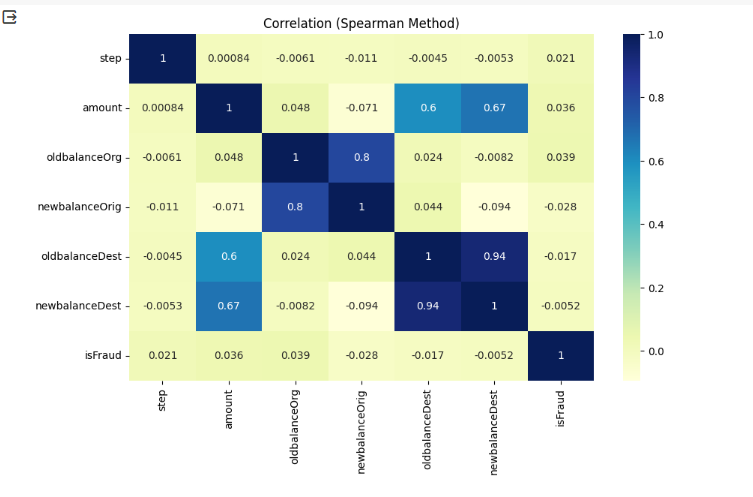
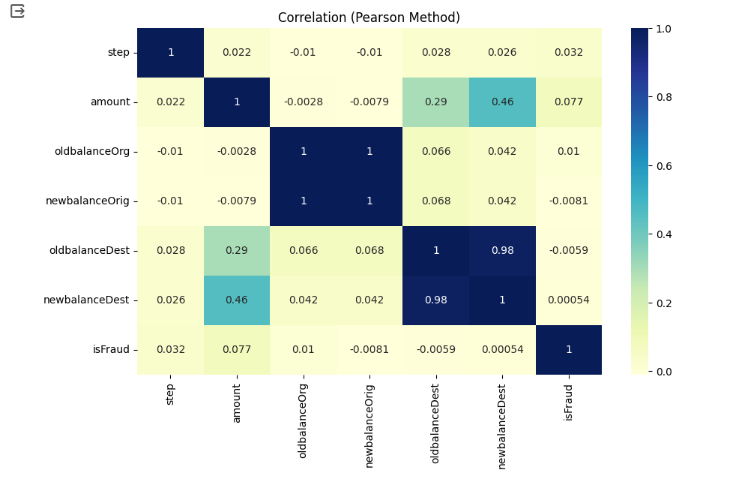
In the exploratory analysis, we refined the dataset to improve credit card fraud detection. We focused on removing columns that seemed irrelevant to the primary goal of identifying fraudulent activities in credit card transactions. As a result, 'isFlaggedFraud,' 'nameOrig,' and 'nameDest' were excluded, creating a more focused dataset for further analyses.

**Histograms:**  


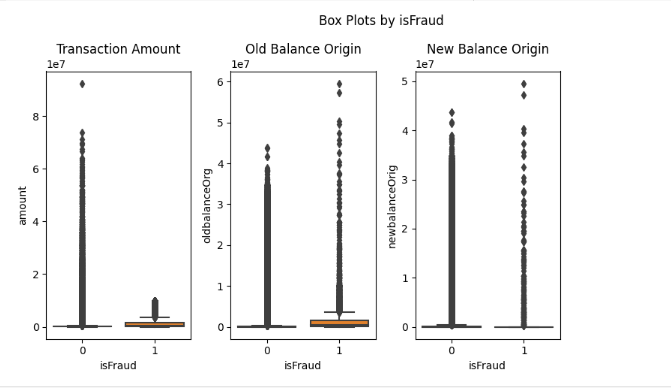
Histograms have illuminated left-skewed data distributions, signaling the

need for normalization to optimize model performance, especially for logistic regression.

**Correlation matrix**



The correlation matrix for both Pearson and Spearman shows that there is a highly positive co-relation between OldBalanceOrig and NewBalanceOrg , and NewBalanceDest and ‘OldBalanceDest’.

**Outliers Visualization** :

To identify possible outliers and grasp the distribution of specific features ('amount,' 'oldbalanceOrg,' and 'newbalanceOrig'), we created box plots, categorized by the 'isFraud' variable. The analysis revealed potential outliers, suggesting a potential imbalance in the distribution between fraudulent and non-fraudulent transactions. Handling outliers required caution due to the scarcity of fraud cases. Removing outliers directly at this point was considered inappropriate, as it might further reduce the already limited number of fraud transactions.

In summary, the exploratory analysis has given valuable insights into our dataset, guiding our path in credit card fraud detection research. We improved focus by removing irrelevant columns, making our dataset more pertinent to fraud detection. Histograms revealed left-skewed data, suggesting a need for normalization.

Moreover, box plots exposed potential outliers, underlining an uneven distribution between fraudulent and non-fraudulent transactions. As part of our feature selection, we're considering removing one of the highly correlated variables to enhance model efficiency.

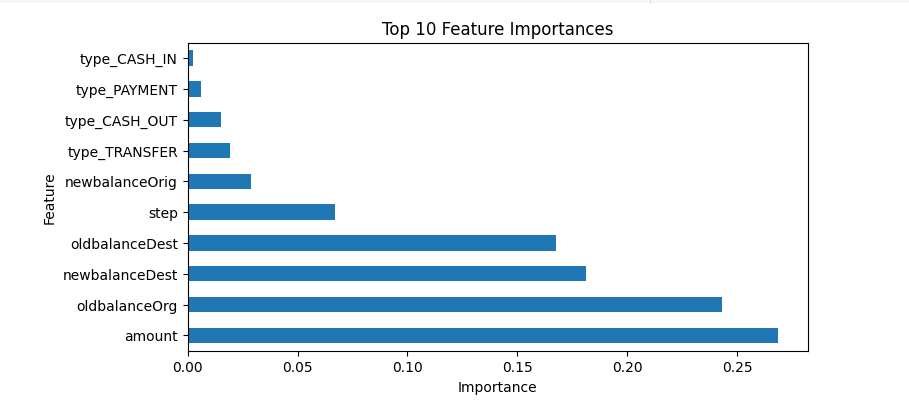
**Dimensionality Reduction**

One-hot encoding is used for converting categorical nominal variable ”type” into a binary format (0 & 1)that is compatible with various machine learning algorithms, ensuring accurate representation and avoiding misinterpretations.

Robust scaling was chosen for its lower sensitivity to outliers, allowing for normalization while maintaining the authenticity of the fraud-related data.

An examination of how fraudulent transactions are spread across various transaction types revealed a higher occurrence of fraud in 'type\_CASH\_OUT' and 'type\_TRANSFER' transactions. Interestingly, no instances of fraudulent activities were observed in the 'type\_DEBIT', 'type\_PAYMENT', and 'type\_CASH\_IN' categories.

From the ANOVA test, it has been found that the p-value (effectively zero) confirms that this variability is highly unlikely to be due to random chance, indicating 'Amount' is a crucial and strong indicator of fraudulent transactions



Random Forest Classifier was used to figure out which features are most important in detecting credit card fraud. It showed a list of 10 influential features.

Now, normally, we might remove features that aren't very important. However, even though 'type\_CASH\_IN' was less important, we chose to keep it. This is because in fraud detection, where finding relevant features is challenging, it's better to include all the information we have.  
From the exploratory data analysis and dimensionality reduction,I decide to delete the ‘NewBalanceOrig’ feature as, ‘NewBalanceOrig’and ‘OldBalanceOrig’ were highly correlated and it seems less important than OldBalanceOrig from the TOP 10 feature importance.

**SPLITTING OF DATASET**

During this phase, the dataset underwent splitting into features (X) and the target variable (y) using a random sampling technique with train: test ratio 80:20. Random sampling entails the random selection of a subset of data points without any specific order. In the realm of credit card fraud detection, this approach proves beneficial as it ensures that the training and test sets encompass a diverse array of transactions.We also verified the proportion of class distribution among Original data, train data and test data.

To counteract the imbalanced distribution within the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed specifically on the training dataset. SMOTE plays a vital role in the realm of credit card fraud detection by creating synthetic instances of the minority class (fraudulent transactions), thereby balancing the dataset. This technique is crucial in preventing the model from developing a bias toward the majority class, consequently enhancing the accuracy of fraud detection. It's important to highlight that SMOTE is exclusively implemented on the training dataset to guarantee that the model encounters untouched data during testing, mirroring real-time scenarios.

Further for the evaluation of models , Time-series Cross validation technique was chosen ,as the datset has “step” column which defines the time , so Time-series Cross validation technique has proved to be good approach in time-dependent data. Along with this, K-fold cross –validation was also considered.

The outcomes of cross-validation revealed varying levels of accuracy across each model. The Random Forest model attained an average accuracy of approximately 95.68%, the Decision Tree model averaged around 96.94%, the Logistic Regression model demonstrated an average accuracy of 80.12% and the Naïve Bayes has an average accuracy of 64.58%.

Apart from high accuracy, Additional metrics such as precision, recall, and F1 score play a pivotal role in a comprehensive evaluation.

Precision indicates the proportion of transactions predicted as fraudulent that are actually fraudulent.

Recall represents the proportion of actual fraudulent transactions that the model correctly identifies.

and the F1 score strikes a balance between precision and recall.Hence, all these metrics has been considered in the performance evaluation of model.

So, Combining random sampling, SMOTE (Synthetic Minority Over-sampling Technique), and time series cross-validation is a good approach, especially for tasks like fraud detection involving imbalanced datasets and temporal dependencies.

**Modeling**

In the modeling phase of the project, four diverse machine learning models were employed to confront the credit card fraud detection challenge. The selected models—Decision Tree, Logistic Regression, Random Forest, and Naive Bayes—each bring distinctive characteristics and strengths to the task.

To commence the modeling process, the preprocessed training data was utilized including random sampling for splitting, applying the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalanced data, and necessary conversions.

**Decision Tree:**

A robust algorithm that partitions the data into subsets based on features, effectively capturing complex relationships.

**Logistic Regression:**

A fundamental algorithm for binary classification, modeling the probability that an instance belongs to a particular category.

**Random Forest:**

An ensemble learning method constructing numerous decision trees during training, enhancing classification accuracy.

**Naive Bayes:**

A probabilistic algorithm based on Bayes' theorem, particularly suitable for classification tasks with relatively simple assumptions.

Each model underwent Time Series Split cross-validation, ensuring an evaluation in a time-dependent manner while preserving the temporal order of transactions. This ensemble of models was carefully chosen to collectively contribute to a comprehensive and effective approach in addressing the credit card fraud detection challenge. The Random Forest model attained an average accuracy of approximately 95.68%, the Decision Tree model averaged around 96.94%, ,the Logistic Regression model demonstrated an average accuracy of 80.12%. And Naïve Bayes has an average accuracy of 64.58%. The acquired results will shape further refinement and optimization in the subsequent stages of the project.

**EVALUATION**

In this phase, all the four models , Decision Tree, Random Forest, Logistic Regression and Naïve Bayes have been assessed on test data .

**Random Forest :** The Random Forest achieves an impressive a**ccurac**y of approximately **99.91%,** indicating that the majority of predictions are correct. A precision of 60.75% suggests that the model predicts a transaction as fraudulent. A recall of 0.88 indicates that the model captures about 88.88% of the actual fraudulent transactions. With an F1 Score of 0.7218, the model demonstrates good overall performance in terms of both precision and recall.

The confusion matrix indicates that the model correctly identified 1440 fraud transactions and 1269974non-fraud transactions. However, it also misclassified 930 non-fraud transactions as fraud and missed 180 actual fraud transactions. An AUC of 99.7% suggests that the Random Forest model has excellent discriminatory power, effectively distinguishing between fraud and non-fraud transactions.

**Decision Tree:**

Decision Tree model achieves an **accuracy** of approximately **99.87%**, indicating that the majority of predictions are correct. A precision of 0.498 suggests that when the model predicts a transaction as fraudulent, it is correct about 49.8% of the time. A recall of 0.886 indicates that the model captures about 88.6% of the actual fraudulent transactions.

With an F1 Score of 0.63, the model demonstrates a reasonable trade-off between precision and recall.

The confusion matrix shows 126,9461 correct non-fraud predictions, 1,443 instances of incorrectly predicting non-fraud as fraud, 184 instances of incorrectly predicting fraud as non-fraud, and 1,436 correct fraud predictions.

The Decision Tree model, with an AUC of 94%, demonstrates good discriminatory ability but may have slightly lower performance compared to the Random Forest.

**Logistic Regression:**

Logistic Regression model displayed lower accuracy at 83.02 %, primarily due to

a precision of 0.49%, which signifies many false positives. The recall, was also

66%. The F1 score, being the harmonic mean, was relatively low at 0.98%. The confusion

matrix indicated that the model struggled with a high number of false positives (215440)

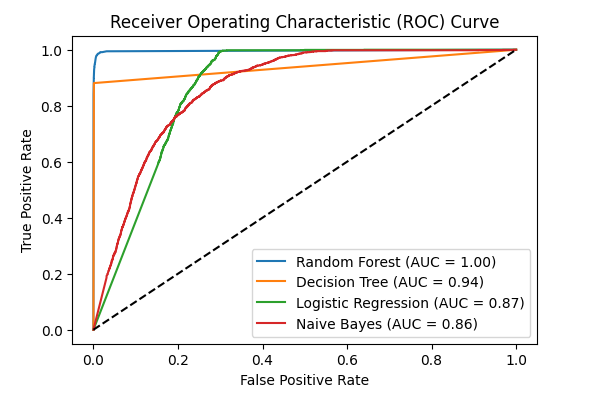
compared to true positives (1072). The AUC for Logistic Regression was 86.68%, reflecting a

moderate ability to distinguish between the two classes.

**Naïve Bayes:**

The Naive Bayes model achieves high recall (93.02%), effectively capturing actual fraud cases, but suffers from extremely low precision (0.31%) and an overall poor F1 Score (0.006), indicating a high rate of false positives. The confusion matrix illustrates a substantial number of false positives (470,835), overshadowing the true positives (1,507). This model may be influenced by the class imbalance, and optimizing for a balance between precision and recall is crucial for its effectiveness in fraud detection.

Random Forest achieves a perfect AUC (1.00), indicating excellent discrimination, while Decision Tree follows closely with a strong AUC (0.94). Logistic Regression and Naive Bayes show fair discrimination with AUC values of 0.87 and 0.86, respectively.  
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 andAUC, Logistic Regression showed limitations, especially in terms of precision.

**LIMITATIONS**

The dataset displayed a pronounced imbalance between classes, posing a significant challenge in the realm of fraud detection. Despite the application of SMOTE, the imbalanced nature persisted. Considering alternative methods like under-sampling or advanced ensemble techniques could bolster the model's capacity to generalize effectively.

The project faced practical limitations due to restricted computing power. Managing a large dataset and computationally demanding models, notably Random Forest, necessitated thoughtful consideration of computational resources. Striking a balance between model complexity and computational efficiency was crucial for feasible execution times. This emphasizes the significance of code optimization, exploring parallel computing options, and considering alternative algorithms that balance accuracy with computational demands.

One significant challenge stemmed from limitations in the dataset's attributes. Despite its extensive transaction volume, the available features were relatively constrained. This restriction hampered the capacity for in-depth feature engineering. The scarcity of relevant features underscored the difficulty in developing a more sophisticated understanding of the factors contributing to credit card fraud.

In future phases of the research, it is imperative to explore avenues for enhancing the dataset by incorporating additional relevant variables. This augmentation should focus on introducing pertinent factors that provide deeper insights into the intricate dynamics of credit card transactions. The aim is to cultivate a more comprehensive dataset, thereby contributing substantially to the development of more effective and robust fraud detection models. This proactive approach ensures that the model is equipped with richer contextual information, potentially improving its accuracy and adaptability in identifying fraudulent activities.

**CONCLUSION**

In conclusion, this study delved into the details of credit card fraud detection, employing machine learning models—Random Forest, Decision Tree, and Logistic Regression. Interestingly, the Decision Tree model outshone the others, showcasing superior accuracy.

A critical finding underscored the significance of dataset balance, and SMOTE emerged as a vital tool in addressing class imbalance. The study's practical implications extend to bolstering real-time fraud detection in financial institutions.

The incorporation of timed cross-validation emphasized the temporal dimension in model evaluation. While the research yielded valuable insights, recognizing inherent limitations, such as dataset attributes and class imbalance, paves the way for future enhancements. Suggestions include richer datasets, advanced techniques beyond SMOTE, improved computing resources, and the collection of more comprehensive temporal details.

This study serves as a stepping stone for future research in credit card fraud detection, emphasizing the continual evolution of methodologies to build effective models.

**REFERENCES**

[1].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>

**[2]** G. Niveditha, K. Abarna, and G. v. Akshaya, “Credit Card Fraud Detection Using Random Forest Algorithm,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 2, pp. 301–306, Apr. 2019,

Available at doi: 10.32628/CSEIT195261.

**[3]** Jain, V., Agrawal, M. and Kumar, A. (2020). Performance analysis of machine learning algorithms in credit cards fraud detection., DOI: 10.1109/ICRITO48877.2020.9197762.

[4]. Kilickaya, Ozlem, Credit Card Fraud Detection: Smote Technique to Improve Model Performance. Available at <http://dx.doi.org/10.2139/ssrn.4412674>

[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] John O. Awoyemi, Mr. Adebayo O. Adetunmbi, and Mr. Samuel A. Oluwadare,2017, Credit card fraud detection using machine learning techniques: A comparative analysis,available at, DOI: 10.1109/ICCNI.2017.8123782.