

# DATA SCIENCE DIGITAL ASSIGNMENT - I

CREDIT CARD FRAUD

DETECTION REPORT

## By TEAM - 7

| SAMRIDDHI MISHRA | 21BCE0586 |
| --- | --- |
| BBM SATHVIK | 21BCE0270 |
| RAHUL AG | 21BCE3036 |
| ARNAV KHARE | 21BCE2092 |
| LIKULEASH | 21BCE2374 |
| ARNAV YADAV | 21BDS0037 |
| ANNANT SHARMA | 21BDS0330 |

**TABLE OF CONTENTS**

| **SERIAL NO.** | **TOPIC** | **PAGE NO.** |
| --- | --- | --- |
| 1. | ABSTRACT | 2 |
| 2. | INTRODUCTION | 3 |
| 3. | UNDERSTANDING CREDIT CARD FRAUD DETECTION | 4 |
| 4. | DISCOVERY PHASE | 5 |
| 5. | DATA PREPARATION | 6 - 11 |
| 6. | MODEL PLANNING | 12 - 13 |
| 7. | MODEL BUILDING | 14 - 15 |
| 8. | COMMUNICATE RESULTS | 16 |
| 9. | OPERATIONALIZE | 17 |
| 10. | CONCLUSION | 18 - 19 |
| 11. | REFERENCES | 20 |

**ABSTRACT**

The credit card has become the most popular payment method for both online and offline transactions. The necessity to create a fraud detection algorithm to precisely identify and stop fraudulent activity arises as a result of both the development of technology and the rise in fraud cases. This paper implements the random forest (RF) algorithm to solve the issue in the hand. A dataset of credit card transactions was used in this study. The main problem when dealing with credit card fraud detection is the imbalanced dataset in which most of the transactions are non-fraud ones. To overcome the problem of the imbalanced dataset, the synthetic minority over-sampling technique (SMOTE) was used. Implementing the hyperparameters technique to enhance the performance of the random forest classifier. The results showed that the RF classifier gained an accuracy of 98% and about 98% of F1-score value, which is promising. We also believe that our model is relatively easy to apply and can overcome the issue of imbalanced data for fraud detection applications.

**INTRODUCTION**

Financial fraud is a serious problem that is only getting worse and has far-reaching effects on the financial sector, businesses, and the government. Fraud is defined as criminal deception done with the intention of making money. Credit card transactions have surged thanks to a high reliance on internet technology. The rate of credit card fraud is rising as credit card transactions take over as the preferred method of payment for both online and offline transactions.

There are two types of credit card fraud: internal and external. While external card fraud entails using a stolen credit card to obtain money through illegal ways, inner card fraud happens as a result of an agreement between cardholders and the bank and involves using a fake identity to commit fraud. Most credit card frauds are external card fraud, which has been the subject of much investigation. Another classification has been made into three categories: classic card-related frauds (application, stolen, account takeover, fake, and counterfeit), frauds involving retailers (merchant collusion and triangulation), and frauds involving the internet (site cloning, credit card generators, and false merchant sites).

Due to their time-consuming nature and ineffectiveness, manual methods of fraud detection have become increasingly impracticable with the introduction of big data. The challenge of credit card fraud, however, has drawn the attention of financial institutions to current computational approaches. One significant way for detecting credit fraud is the use of data mining techniques. The technique of separating fraudulent transactions into two categories: legitimate and fraudulent transactions is known as credit card fraud detection. Machine learning has many techniques to solve this problem including but not limited to Artificial Neural Networks (ANNs), Naive Bayesian Classifier, support vector machines (SVMs), decision trees, random forests, k-nearest neighbors, logistic regression, intelligent decision engines (IDE), meta-learning strategy.

**UNDERSTANDING CREDIT CARD FRAUD DETECTION**

**What is Credit Card fraud Detection?**

Credit card fraud detection is the process of identifying and preventing fraudulent transactions before they can cause financial damage. It involves using advanced algorithms and machine learning techniques to analyze transaction data in real-time and detect patterns that may indicate fraudulent activity.

**KEY FEATURES:**

1. Enormous Data Processing : Credit card companies handle massive data daily, requiring fast fraud detection.

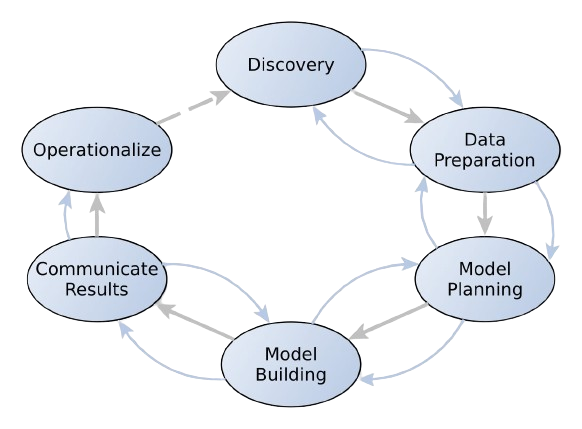
2. Imbalanced Data : About 99.8% of transactions are legitimate, making fraud detection challenging.

3. Data Privacy : Protecting user privacy is vital; dimensionality reduction techniques help maintain privacy.

4. Misclassified Data : Improving detection accuracy requires addressing misclassified transactions.

5. Adaptive Techniques : Scammers evolve tactics; simple, interpretable models enable quick deployment of new detection methods.

The data analytics lifecycle of a credit card fraud detection system encompasses several steps outlined below, which are further elaborated upon in the report.



# **EXPLORING THE ROLE OF CREDIT CARD FRAUD DETECTION IN THE DATA DISCOVERY PHASE**

1. **Learning the Business Domain:**

Grasping the intricacies of credit card transactions, fraud detection regulations, and prevalent fraud schemes.

1. **Assessing Resource Availability:**

We assessed the resources available for our project, including access to data, computational resources, software tools, and expertise. We considered whether we could obtain real-world credit card transaction data or use simulated data for our analysis. Additionally, we identified relevant libraries or frameworks for data analysis and machine learning that we could leverage.

1. **Framing the Problem:**

Defining specific objectives, constraints, and success criteria for the fraud detection system. Considering factors such as fraud detection accuracy, false positive rates, regulatory compliance, and customer experience.

***THE PROBLEM STATEMENT:***

* *To Develop an effective fraud detection system exclusively for credit card transactions, aiming to identify and prevent unauthorized activities while maintaining the integrity of legitimate transactions and ensuring customer satisfaction.*

1. **Identifying Key Stakeholders:**

Engaging with financial institutions, payment processors, merchants, regulatory bodies, and cardholders to understand their perspectives and requirements.

1. **Interviewing the Analytics Sponsor:**

Conducting interviews with senior executives or decision-makers to gain insights into strategic goals, risk tolerance, and expectations for fraud detection.

1. **Finding Potential Data Sources**:

Identifying and exploring various data sources such as transaction logs, user profiles, merchant information, and external databases to gather relevant information for fraud detection.

1. **Developing the Initial Hypothesis:**

Formulating educated guesses about potential fraud patterns and behaviors based on domain knowledge. Hypothesizing factors such as transaction timing, amounts, merchant categories, and user behavior.

**2.** **OPTIMIZING DATA FOR FRAUD DETECTION:** **STRATEGIES IN DATA PREPARATION**

**STEPS INVOLVED:**

1. **Sandbox Setup:**

* **Objective:** Create workspace for modeling.
* **Actions**: Establish an isolated environment, install essential tools.
* **Example:** Setup Python environment for analysis

1. **ETLT Process:**

* **Objective**: Extract, transform, load, and validate data.
* **Actions**: Extract transaction data, clean, transform, test quality.
* **Example**: Extract records, clean timestamps, ensure integrity.

1. **Data Understanding:**

* **Objective**: Understand data structure and context.
* **Actions**: Explore schema, identify key features, learn limitations.
* **Example**: Study transaction dataset for attributes.

1. **Data Cleaning:**

* **Objective:** Address inconsistencies and missing values.
* **Actions:** Remove duplicates, impute missing values, correct errors.
* **Example:** Remove duplicate transactions, fill missing timestamps.

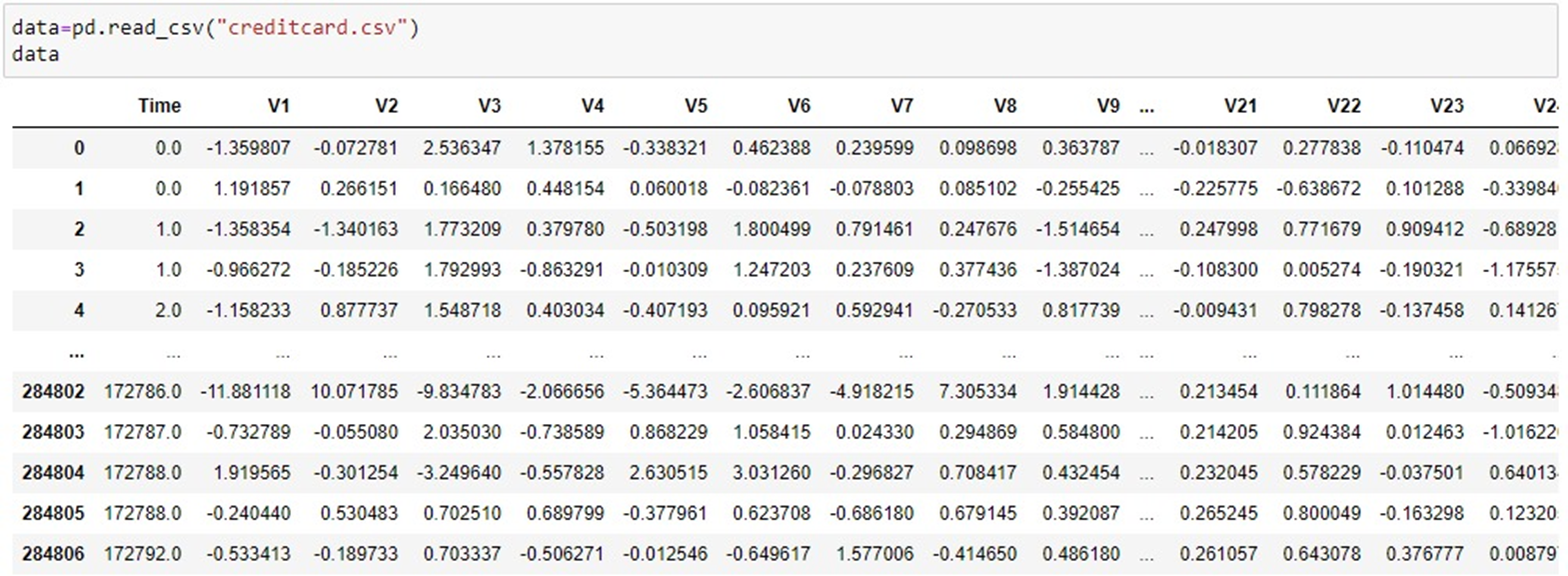
1. **Survey & Visualization:**

* **Objective:** Explore distribution, detect anomalies.
* **Actions:** Plot histograms, analyze stats, detect outliers.
* **Example:** Visualize transaction amounts over time.

1. **Common Tools:**

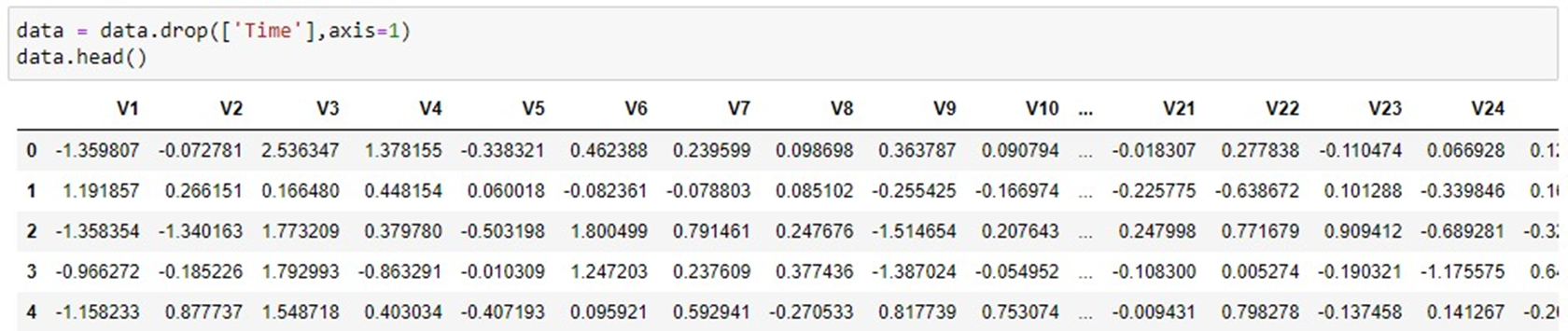
* **Objective**: Utilize tools for streamlined processing.
* **Actions**: Use Pandas for filtering and cleaning data.
* **Examples:** Pandas, NumPy, Seaborn/Matplotlib, SQL.

**Understanding the Data Set:**

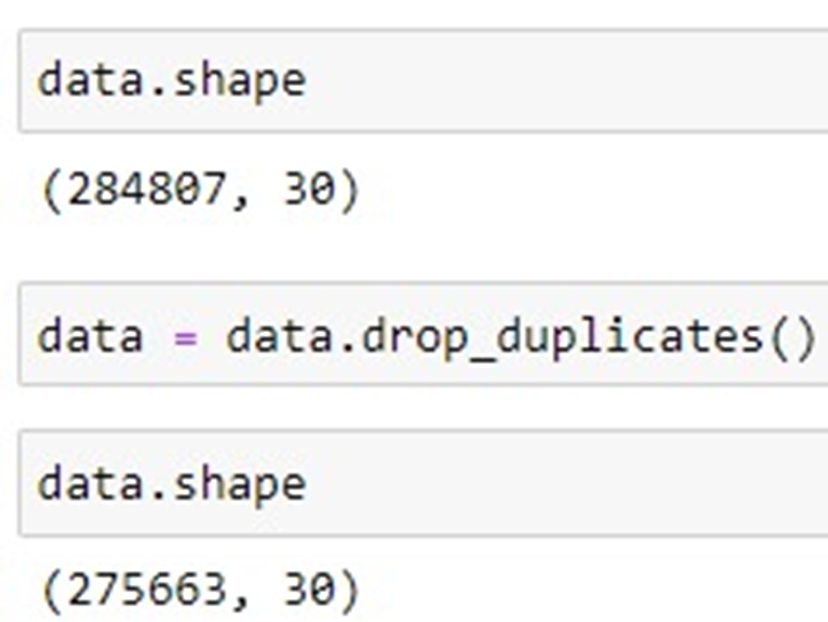


**Data Transforming:**



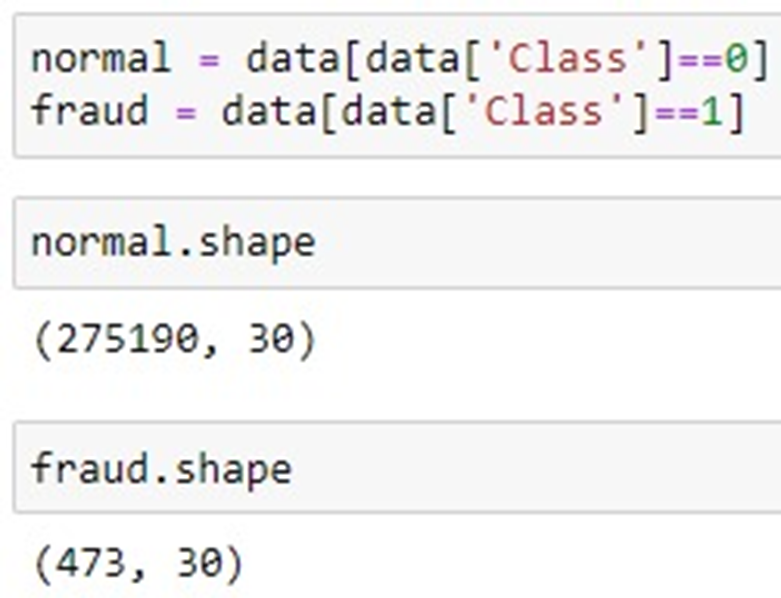


**Removing Duplicate Entries:**

****

**Identifying Normal and Fraudulent Transactions:**

Upon inspecting the dataset, we find that the majority of transactions are non-fraudulent (Class 0), with only a small proportion being fraudulent (Class 1). This indicates a class imbalance, which is common in fraud detection datasets. Dealing with this imbalance is crucial for developing accurate predictive models.



**Final Data Review:**

****

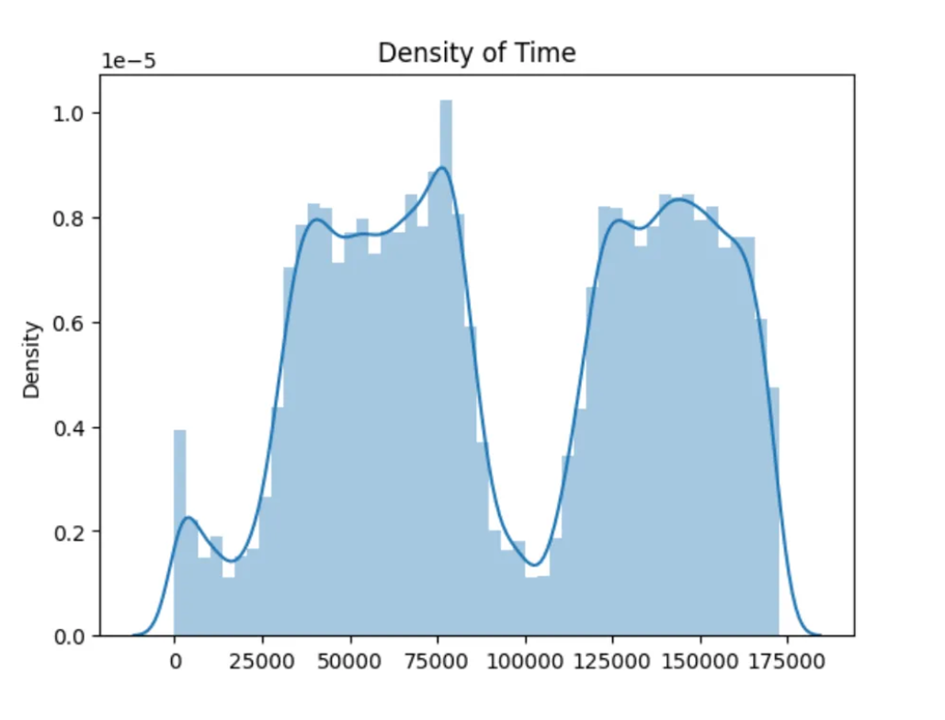
****

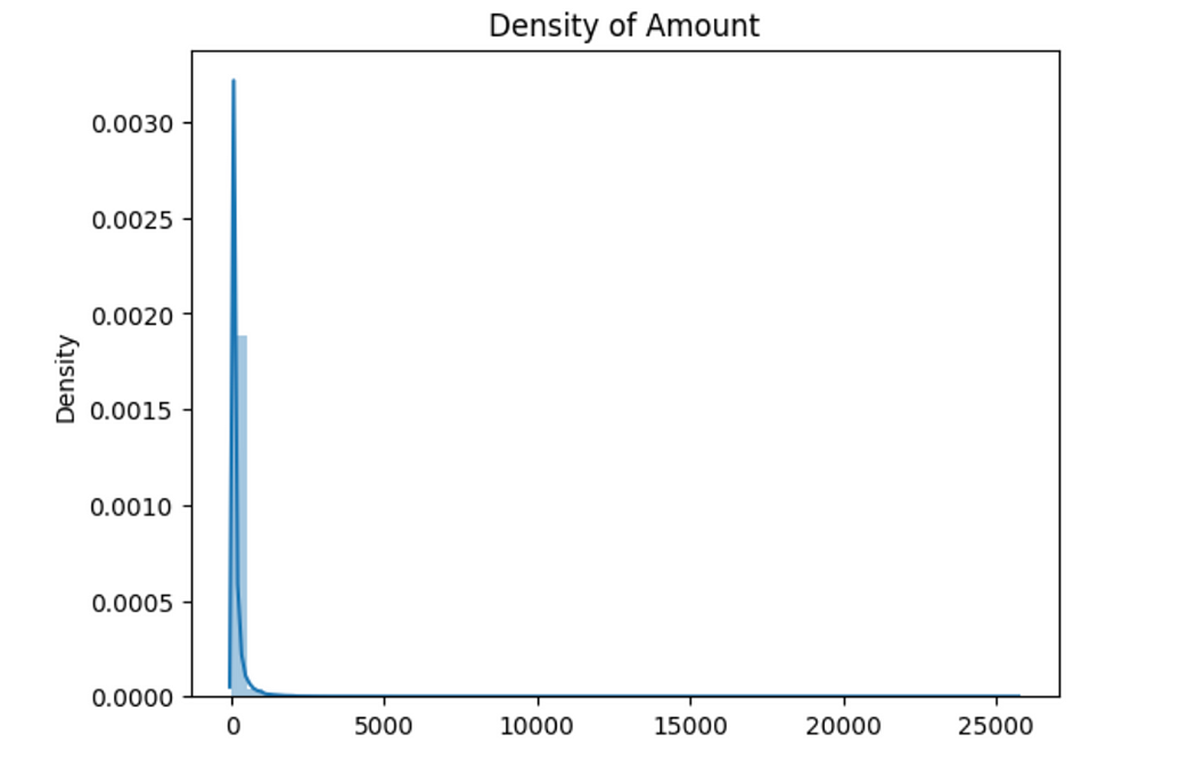
**Data Visualization**

1. **Scatter Plot**

****

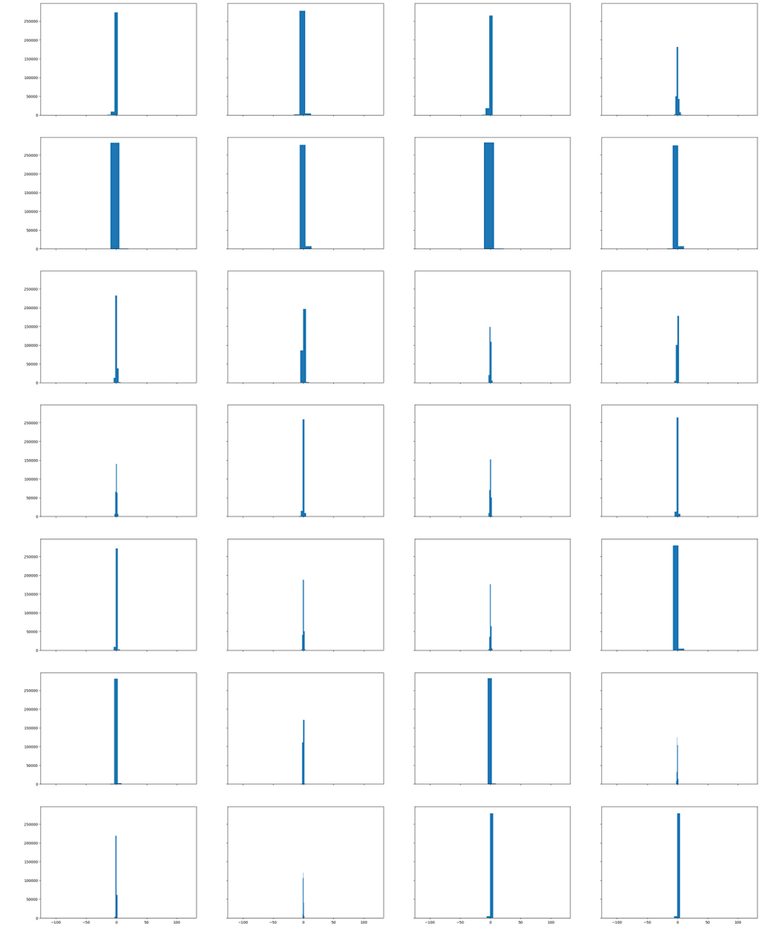
1. **Density Plot**

****

****

The density plot of the “Time” feature provides us with insights into the distribution of transaction times. By examining the plot, we can identify the concentration of transactions at different time intervals. This information can help detect potential anomalies or fraudulent activities occurring during specific periods. Analyzing the density plot of the "Amount" feature in a credit card fraud detection dataset reveals insights into transaction value distribution. This visualization helps identify typical transaction ranges and detect potential anomalies, aiding in the development of robust fraud detection models and identification of suspicious activities.

1. **Histogram**

****

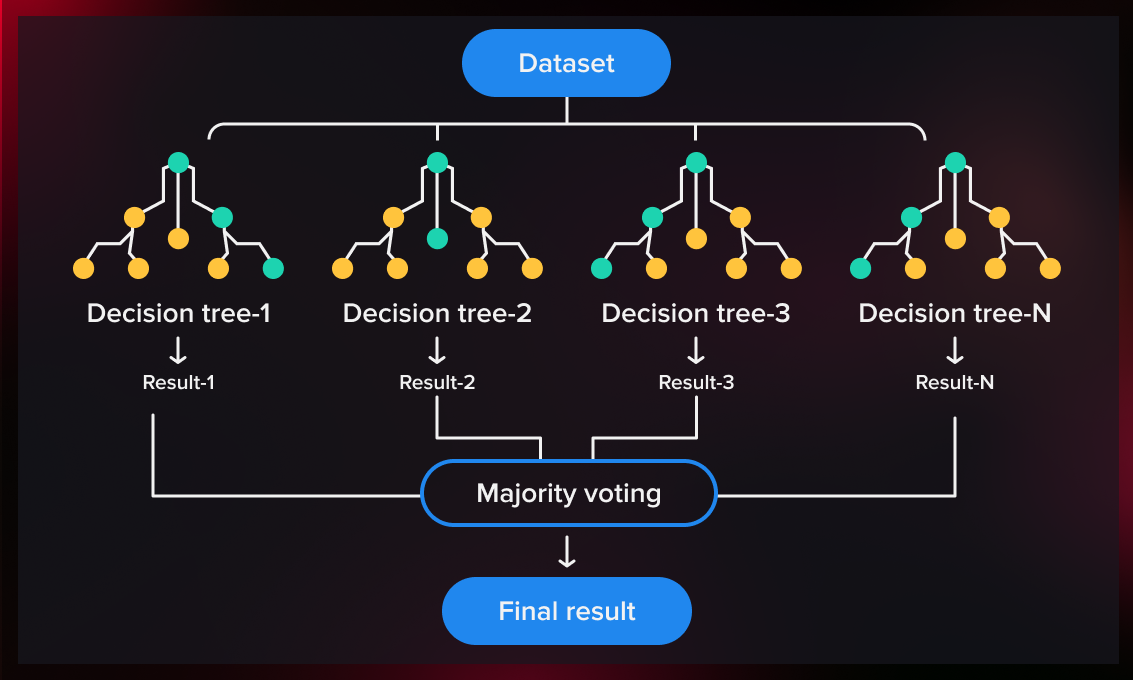
Histograms visually represent feature distributions, offering insights into shape, central tendency, and spread. Patterns like symmetry, skewness, or multimodality indicate characteristics or outliers. Analyzing these in a fraud detection dataset unveils data nature, aiding accurate model development.

# **3. POSSIBILITIES OF MODELS INVOLVED IN CREDIT CARD FRAUD DETECTION : MODEL SELECTION**

1. **Logistic Regression:**
   * Logistic regression is a statistical method used for binary classification, which means predicting the probability of occurrence of an event or outcome.
   * Simple and interpretable.
   * Dependent variable is dichotomous(binary).
   * explain the relationship between one dependent binary variable and one or more independent variables.
   * It assumes a linear relationship between the independent variables and the log-odds of the outcome.
   * Fast Training and Prediction: Logistic regression is computationally efficient and can handle large datasets with relatively low memory and CPU requirements. Training and prediction times are typically faster compared to more complex models like neural networks.
   * Example: Good starting point for its simplicity.

**2) Random Forest (RF):**

* + Handles non-linear relationships.
  + Random Forest is a classification algorithm that is comprised of many Decision Trees.
  + Each tree has nodes with conditions, which define the final decision based on the highest value.
  + ­It is like a group decision-making team in machine learning.
  + ­It combines the opinions of many “trees” (individual models) to make better predictions of models.
  + ­It tackles both classification and regression problems effectively.
  + It has the ability to handle complex Datasets & Robust against overfitting.
  + It can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification.
  + Example: Suitable for imbalanced data, captures complex patterns.



**Data Instances involved in the credit card fraud detection:**

* Data instances for credit card fraud detection typically consist of individual credit card transactions, each represented as a data point.
* Each instance contains various features (attributes) that describe the transaction, and a label indicating whether the transaction is fraudulent or legitimate.
* Amount with transaction location.
* Amount with Previous transaction Location.
* Card Holder’s Information.
* class with merchant\_id.
* Type of Transaction.
* Fraudulent or legitimate: A binary label indicating whether the transaction is fraudulent (1) or legitimate (0).he transaction is fraudulent (1) or legitimate (0).

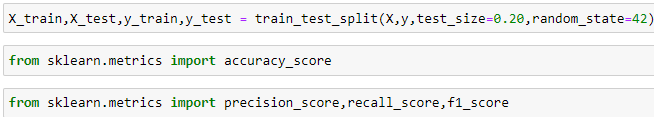
**Key Features of credit card fraud detection:**

* Time since last transaction.
* Frequency of transactions.
* Transaction amount relative to cardholder's typical spending pattern.

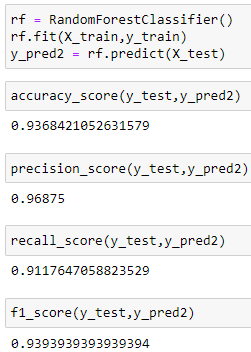
According to the Correlation matrix these three features are the most dependent variables to consider whether the fraud has been detected or not.

**4. SHAPING** **THE SOLUTION: THE MODEL BUILDING PROCESS**

**Training the model using Random Forest Algorithm:**



**Implementing the model on test data set:**



**Software used:**

For this project, we utilized Jupyter Notebook, an interactive web-based application that facilitates seamless code execution, documentation, and visualization. Within Jupyter Notebook, we imported essential libraries in Python, leveraging their functionalities for tasks such as data manipulation, statistical analysis, and machine learning. We employed this environment to meticulously craft, train, and fine-tune our model, iterating through various algorithms and parameters to optimize performance. Finally, we rigorously tested our model on the dataset, assessing its accuracy, robustness, and generalization capabilities to ensure its effectiveness in real-world scenarios.

**WHY RANDOM FOREST?**

1. **Handles Imbalanced Data:**

Credit card fraud data is imbalanced, with far more legitimate transactions than fraudulent ones. Random Forest performs well even with imbalanced datasets because it trains multiple decision trees and averages their predictions.

1. **Robust to Overfitting:**

Random Forests are less prone to overfitting the training data compared to some other models like decision trees. This is because each tree uses a random subset of features, reducing the chance of memorizing irrelevant patterns.

1. **Interpretability:**

Unlike complex models like neural networks, Random Forest allows you to understand the features that contribute most to fraud detection. This can be valuable for improving future models and explaining decisions.

1. **Accuracy:**

Random Forest often achieves high accuracy in fraud detection tasks. It aggregates predictions from multiple trees, leading to a more robust and generalizable model

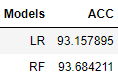
**RANDOM FOREST V/S OTHER MODELS**

* **Logistic Regression:** While interpretable and good for binary classification, it might struggle with complex fraud patterns and imbalanced data.
* **Gradient Boosting:** (e.g., XGBoost) Can be very powerful but requires careful hyperparameter tuning and might be less interpretable than Random Forest.
* **Support Vector Machine (SVM):** Effective for some fraud detection tasks, but can be computationally expensive for large datasets and might be less robust to imbalanced data.
* **Neural Networks:** Can be very powerful, especially for complex patterns, but require significant data and computational resources. They also lack interpretability.
* **Ensemble Methods (including Random Forest):** These can be even more powerful than individual models like Random Forest by combining multiple models. However, interpretability can become a challenge.
* **Anomaly Detection Models:** These are good for identifying outliers, but might not be ideal for classification tasks like fraud detection where specific patterns need to be learned.

1. **EMPOWERING DECISION-MAKING: COMMUNICATING FRAUD DETECTION FINDING**

The team succeeded in making a credit card fraud detection model. The results were statistically significant and valid. The effectiveness of different models in detecting credit card fraud, highlighting the best-performing model(s) based on evaluation metrics

We had 2 models to start with. These were logistic regression and random forest. We then went on to choose random forest as its real world implications are better and also is more efficient to use for bigger data.

****

**Insights into the importance of various features in identifying fraudulent transactions:**

* Transaction Amount: Unusual transaction amounts, especially those significantly higher or lower than the cardholder's typical spending patterns
* Transaction Frequency and Timing: Unusual patterns in transaction frequency or timing, such as a sudden increase in transactions or transactions occurring at odd hours
* Geolocation: Monitoring the geographic location of transactions can help detect fraudulent activities

**Any patterns or trends observed in fraudulent activities, such as time of day, transaction amount, or geographical location**

* Time of Day: Fraudulent activities often occur during off-peak hours when monitoring systems may be less active.
* Transaction Amount: Fraudulent transactions often involve unusually high amounts, which may exceed the cardholder's typical spending patterns.
* Geographical Location: Fraudulent transactions may occur in locations that are different from the cardholder's usual spending areas.

**Recommendations for improving fraud detection accuracy or strategies for preventing fraudulent activities based on the analysis results**

Enhance anomaly detection techniques, implement real-time monitoring, utilize advanced machine learning algorithms, incorporate behavioral analytics, implement multi-layered authentication, enhance data sharing and collaboration and educate cardholders.

# **6**. **OPERATIONALIZE**

**Key Findings:**

1. **Critical Predictors Identified:** Transaction timing, amounts, and user behavior are key fraud indicators.
2. **Fraud Patterns Discovered:** Fraud peaks at specific times and locations, revealing distinct trends.
3. **Improvement Recommendations:** Strengthen anomaly detection, implement real-time monitoring, and deploy advanced ML algorithms.

**Deployment Considerations:**

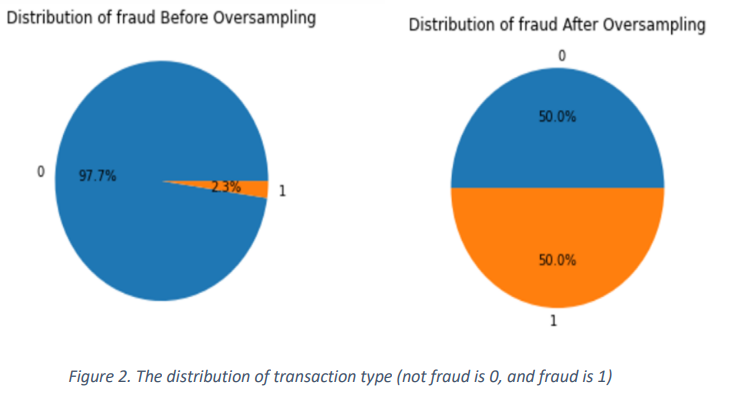
1. While various aspects of credit card fraud detection have been explored, deployment strategies were not thoroughly discussed.
2. Deployment involves integrating the developed model into production systems, considering scalability, reliability, and integration with existing infrastructure.
3. Effective deployment requires collaboration with IT teams, ensuring seamless integration and adherence to regulatory standards

**Key Features:**

1. **Model Deployment:** Implement the selected Random Forest model into a real-time environment. This involves integrating the model into the existing credit card transaction processing system of the financial institution.
2. **Privacy Preservation:** Ensure that the model training process respects user privacy. Employ techniques like differential privacy or federated learning to train the model without exposing sensitive user information. Must not violate RBI guidelines for customer Data.
3. **Collaboration with Financial Institutions:** Foster collaboration between financial institutions to collectively benefit from the fraud detection system. Share insights, methodologies, and best practices to improve fraud detection across the industry. Eg:- Mastercard and Visa have collaboration with transaction data provided by the bank for fraud detection.
4. **Compliance and Regulation:** Address regulatory compliance requirements and adhere to industry standards for fraud detection and data privacy. Work closely with regulatory bodies to ensure that the deployed system meets all necessary legal and ethical guidelines. Must follow all RBI guidelines.
5. **Security Measures:** Implement robust security measures to safeguard the deployed system against potential cyber threats and attacks. Eg:-Tokenization replaces sensitive credit card data with unique tokens. These tokens are meaningless to attackers.
6. **Training and Education:** Provide training and education to stakeholders, including bank personnel, on the use and benefits of the fraud detection system. Build confidence among users and decision-makers regarding the effectiveness and reliability of the deployed technology

**CONCLUSION**

This study implemented an enhanced random forest algorithm using hyperparameters to find the best values to run the classifier and gain the highest possible accuracy. We performed the SMOTE oversampling technique to overcome the imbalanced dataset issue. The result showed that the classifier was able to predict fraudulent transactions with an accuracy of 98%. We evaluate our results using different metrics such as recall, precision, F1 score, and the ROC curve to improve that our classifier is not biased because of the imbalanced dataset and that it’s able to detect both classes (fraud, not fraud) with a high accuracy level. Results showed that our model, due to its relatively straightforward application, is well-suited to addressing the challenge of imbalanced data in fraud detection.



**1. Basic Understanding:**

Credit card fraud detection is the process of identifying and preventing fraudulent transactions by using advanced algorithms and machine learning techniques. Credit card companies handle massive amounts of data on a daily basis, and to top it off about 99.8% of transactions are legitimate, making fraud detection even more challenging.

**2. Phases :**

In the above example also we go through the 6 phases of Data Analytics Lifestyle which are -

· Data Discovery: Involves Grasping credit card fraud nuances, assessing resources, framing objectives, engaging stakeholders, finding potential data source etc

· Data preparation: Includes setting up sandbox, ETLT process, data understanding, data cleaning and survey.

· Model Planning: Herein we decide which model to choose and in our case we have taken the Random forest model

· Model Development: In this phase we implement the random forest model

· Communicating Results: The outcome is that our team has successfully implemented the model

· Operationalize: This phase mainly deals with model deployment for real world application.

**3. Random Forest Model :**

Random forest algorithm is basically a supervised machine learning algorithm that works using multiple decision trees and since it is an ensemble algorithm , it works using multiples of itself resulting in a forest of trees

**4. Implementation and result :**

The use of Random Forest model for credit card fraud detection was successful as its ability to handle imbalanced data, prevent overfitting, and providing crucial interpretability came in very handy as it very effectively Identified critical predictors, including transaction timing, amounts, user behavior , geographic location and transaction frequency to offer actionable insights , which would show potential fraudulent transactions.

**5. Key Advantages of Random Forest Model :**

· The random forest algorithm is not biased, since there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd" therefore, the overall biasedness of the algorithm is reduced.

· This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.

· The random forest algorithm works well when you have both categorical and numerical features and also the algorithm performs better with a larger number of training data

· The Random Forest Algorithm is among the best performers and is known to have fantastic accuracy and precision score and also is pretty efficient and fast.

**REFERENCES**

<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

<https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/>

<https://www.infosysbpm.com/blogs/bpm-analytics/machine-learning-for-credit-card-fraud-detection.html>

<https://www.infoq.com/articles/fraud-detection-random-forest/>