# Exploring the Influence of Transaction Amounts on Fraudulent Activities in Credit Card Transactions

**Author:**

Name- Chandrika Bhargavi Achanta

# Abstract

This research investigates the relationship between transaction amounts and the likelihood of fraudulent activities in credit card transactions using the Credit Card Fraud Detection Dataset 2023. With the prevalence of credit card fraud posing significant threats to financial security, this study employs statistical analysis and advanced machine learning models to discern patterns and for predicting fraud. Initial findings from a subset of data indicated minimal correlation between transaction amounts and fraud incidence. Nevertheless, upon analysing an extended data sample extensively, more distinctive trends emerged, with machine learning models such as Logistic Regression and Gradient Boosting Machines coming to the force as highly effective in the identification of fraudulent transactions across various bands of activity.

The study underscored the crucial significance of data dimensionality in advancing the model accuracy and reliability. It was found that the large-scale dataset allows the model to learn the fraudulent behaviour better, which improves the precision and recall metrics significantly. The paper discussed addressing the problem of class imbalance and data quality, proposing some methods to enhance the model robustness and practicability.

These findings contribute to the continuous enhancement of fraud detection systems, which offer insights that can lead to better targeted and more effective fraud prevention strategies in financial institutions. This research leads us to further study of the integration of these models into real-time systems as well as to research on their effectiveness in different transaction environments.

**Keywords:** Credit Card Fraud, Transaction Amounts, Fraud Detection, Logistic Regression, Gradient Boosting Machines, Machine Learning, Data Analysis.

.

# Table of Contents

# 1. Introduction

## 1.1 Background

Credit card fraud is still an epidemic all around the world and is in constant evolution due to technological advances. It is about unauthorized use of accounts to steal funds and get personal information as payback because of the popularity and the dependability of digital payments method. Since the first digital transaction are being tracked, fraud incidents tend to increase, because the most of online and mobile transactions shift the transactions criteria. Because of that transformation the identification and deterrence of fraud transactions is the primary concern of the financial services.

## 1.2 Problem Statement

Despite advances in fraud detection technologies, financial losses from credit card fraud continue to grow up, with billions of dollars churned out annually. The challenge is persistent mainly because of the highly sophisticated tactics used by fraudsters and the reactive nature of many fraud detection systems. There is a critical requirement for the proactive detection tools to not only react to the fraud activities but predict and prevent them before they happen.

## 1.3 Research Questions

This research is guided by two primary questions:

1. Is there a correlation between transaction amount and the likelihood of a transaction being fraudulent in the Credit Card Fraud Detection Dataset 2023?
2. Can models be tailored to specifically identify high-risk transactions based on amount thresholds, improving the precision of fraud detection efforts?

## 1.4 Significance of the Research

The significance of this research is that it might improve the knowledge of the relationship between transaction size and fraud frequency. By detecting definite patterns and a specific threshold for increased risk, this project will create a collection of warning standards of predictive models that enhance the accuracy and effectiveness of the fraud detecting system. This work have a lot of importance for financial activities because of established guidance to develop a set of strategies that prioritize the high-risk transactions to prevent the fraud rate and for reducing the damage as well.

## 1.5 Contribution to Knowledge

The study contributes to the existing literature by applying advanced data analysis and machine learning techniques to a large-scale, real-world dataset, providing empirical evidence on the effectiveness of these methods in fraud detection. The findings are expected to offer both theoretical and practical insights, advancing the field of financial cybersecurity. Moreover, by enhancing the predictive capabilities of fraud detection systems, this research supports the broader aim of safeguarding consumer transactions and fostering trust in digital financial services.

### 1.6 Potential Implications and Applications

Practically, the outcomes of this research could be integrated into real-time fraud monitoring systems, enhancing their ability to detect and prevent fraudulent transactions. Theoretically, the project expands on the application of machine learning in financial fraud detection, potentially inspiring further studies and technological innovations in the field. This work could also inform policy decisions regarding regulatory standards and practices for transaction security in the financial sector.

## 2. Literature Review

### 2.1 Introduction

In the following research, the relationship among transaction amount as well as credit card fraud being explored with the help of Credit Card Fraud Detection Dataset 2023. Statistical analysis as well as ML models that are logistic regression as well as random forest algorithm is being applied for understanding that high spending relates with improved fraud risk. The segmenting transaction in diverse amount groups following study aims to improve strategies to detect fraud. The primary goal of research is to assist institutions of finance to provide effective resources which focus on transactions having high fraud risk and improve security measures.

### 2.2 Related Work

Credit card fraud developing issues for businesses all around the world and needing the researchers for exploring different applications used by the fraudsters. Various fraudsters using different tactics include authorized account usage, identifying theft, as well as misrepresentation makes it important for understanding the scenario of credit card fraud. Conflicting with the popular belief that merchants mostly suffer with the significant risks which cardholder facing losses of product, chargeback fees, as well as possible account closing. Researchers mostly suggest that the internet transactions mainly “card not present” scene leads to high risk because of absence of verification method. Fraudsters allure internet lies in limited ability for performing traditional check and making hotspot for fraud of credit card [1]. Advancement in recent technology mitigate fraud in following scenario. Different studies showing substantial loss because of the fraud of credit card and projections are reaching billions and needed efficient measures to detect fraud. To understand techniques of fraud like application fraud, stolen cards, counterfeit cards the user need to provide important understandings for battling following crimes. Following literature explains importance of developing technology as well as strategic risk management for staying ahead to protect credit card fraud [10].

Review of techniques of fraud detection within transaction of credit card is important because of rising examples of fraud of credit card within recent years. The credit card becomes a mode of payment and to detect fraud becomes important [8]. Detection of fraudulent activities effectively is important to reduce economic losses. Usage of modern technologies which are data mining, ML, Sequence alignment, and AI introduced to increase the efficiency. The fraud of credit card includes to obtain money by illegal systems and has become significant threat in businesses. The development of technology has led to increase frauds and the detection is challenging risk. Following research explains importance in developing models to detect credit card fraud and recognise increase of fraud due to advancement of technology. Credit card complexities mainly range from a simple threat to the online fraud in which cardholder being present and require advance systems to detect fraud. Limited availability of transaction data and techniques like decision trees, neural network, and case-based reasoning used for creating efficient models to detect fraud. Following review explains the requirement of innovative approach for handling the developing challenges about credit card fraud within advanced technology [2].

Following research paper mainly focus to battle with credit card fraud and issues with digital age. With surge within digital as well as plastic money use the fraudsters mainly develop new ways for causing financial losses. Following research explains about systems to detect fraud and capable to identify as well as accurately detect fraud [6]. The research investigates into different types of fraud related with credit card and includes application fraud, scenario where card-not-present, counterfeit fraud, etc. for addressing following challenges the ongoing investigation explains various techniques to detect fraud. Some of them are Support Vector Machine (SVM), Hidden Markov Model, Fuzzy Logic Based System, and Decision Trees and provides complete study based upon quantitative measures which are accuracy, detection rate, as well as false alarm rate. The literature explains perpetual struggle against fraud of credit card as well as need for advance models. Fraud types which range from application to the mail non-receipt card fraud mainly explains the complexities which are included. The following research contribute towards summarizing techniques, to review fraud detection as well as conduct parametric comparison to select effective techniques. Author concludes by providing drawbacks of models as well as providing solutions for further improvement within credit card fraud detection [4].

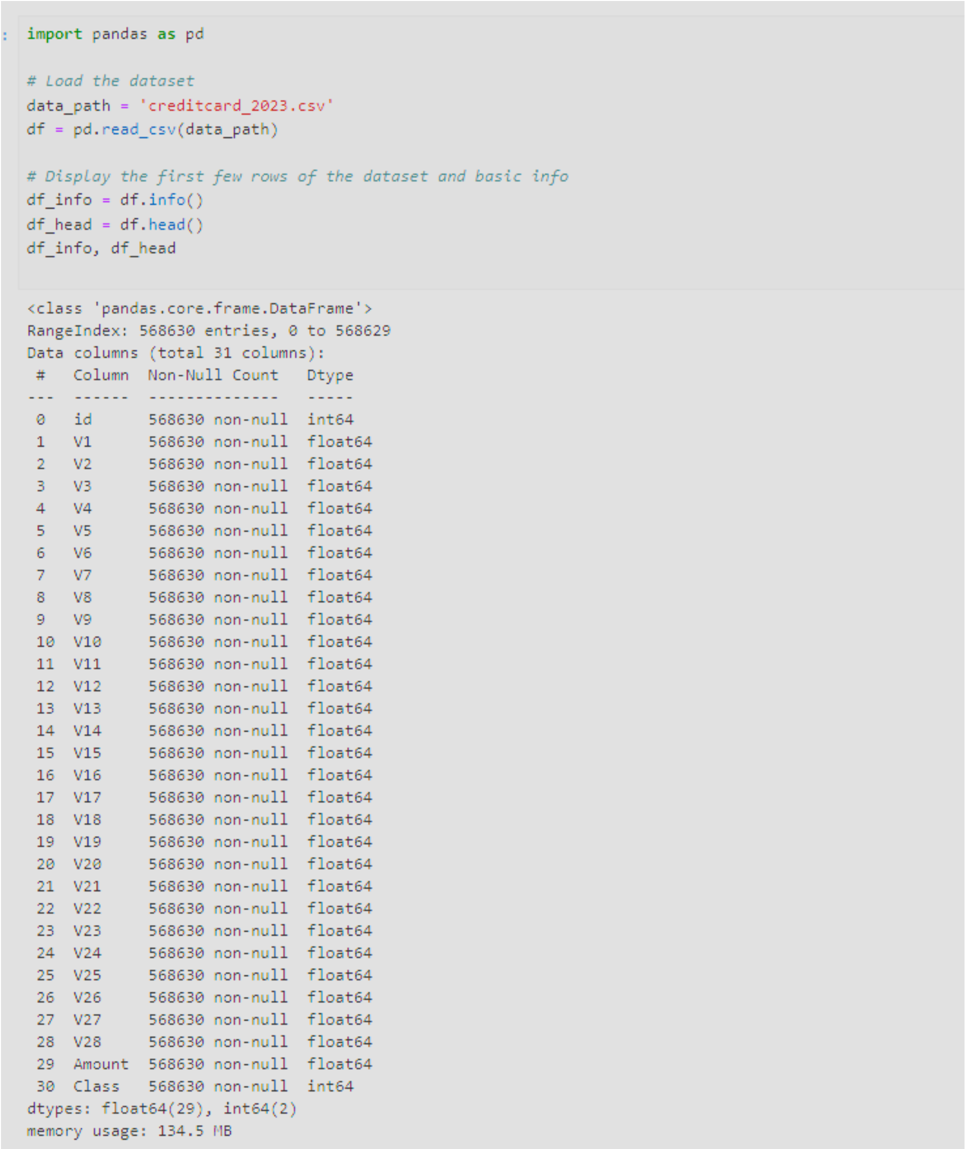
Following research paper explains about growing issues about credit card fraud at the time of online transactions which are causing substantial global losses. Authors contradicting on following illegal activities by data mining as well as techniques of machine learning. The fraud of credit card being categorised into online, and frauds related to merchant, led to significant threat. The research highlight flow in the losses of credit card fraud is reaching about $16.31 billion in 2020 and being estimated to exceed about $35 billion in year 2028. To fight with the fraud following study uses techniques of data mining for studying patterns of transaction to differentiate between suspicious as well as non-suspicious patterns. The techniques of machine learning, mainly Bayesian network classifiers which are Naïve Bayes, K2, TAN, Logistics, and J48, being utilised for predicting automation [9]. Following research explain about supervised classification improving techniques of data pre-processing like normalization as well as principal component analysis. Classifiers display accuracy for exceptional 95% after pre-processing data set being compared with results. Following literature review explain significance to use data mining as well as machine learning techniques for detecting fraud of credit card [5].

Author have explored applications of data mining mainly neural network within the credit card fraud detection systems. The fraud of credit card mainly led to significant threat and the study uses unsupervised architecture of neural network mainly Self-Organizing Map Neural Network (SOMNN) for classifying transactions in clusters like low, high, risk, and high-risk. Following system uses Receiver-Operating Curve (ROC) for detection of credit card fraud to achieve about 95% of detection of the fraud cases without the false alarm and outperform statistical models. Research explains strategic necessity of timely information about fraudulent activities belonging to banking industry in which data mining acts as important tool for extracting business understanding from vast dataset. The detection of credit card fraud includes to categorise transaction into fraudulent classes and three important categories are traditional card-related frauds, merchant-related frauds, and internet frauds. Following study introduce data mining as process which uses different tools for discovering pattern and focus upon neural network because of ability of adaptability [7]. Following research mainly explains about the model which is Credit Card Fraud Watch (CCFW) that uses self-organizing artificial neural networks and transaction rules to detect breaches of transaction policy in real-time. The system to detect CCF work efficiently in banking environment to use advantages of neural networks to learn from previous data and improve results as per the time. Following study explains importance to develop methodologies of intelligent fraud detection within operational system [3].

## 3. Methodology

### 3.1 Dataset Acquisition

The dataset utilized for this research was sourced from Kaggle’s Top rated Credit Card Fraud Detection Dataset 2023, a comprehensive compilation designed to facilitate the development and testing of predictive models aimed at detecting fraudulent transactions. The dataset encompasses over 568,630 transactions, each characterized by 31 distinct features, including transactional details and anonymized user information. Key features include transaction amount, time, and 29 other numerical inputs derived via Principal Component Analysis (PCA) to ensure privacy and data protection.



```
import pandas as pd

# Load the dataset
data_path = 'creditcard_2023.csv'
df = pd.read_csv(data_path)

# Display the first few rows of the dataset and basic info
df_info = df.info()
df_head = df.head()
df_info, df_head
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 568630 entries, 0 to 568629
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   id      568630 non-null  int64  
 1   V1      568630 non-null  float64
 2   V2      568630 non-null  float64
 3   V3      568630 non-null  float64
 4   V4      568630 non-null  float64
 5   V5      568630 non-null  float64
 6   V6      568630 non-null  float64
 7   V7      568630 non-null  float64
 8   V8      568630 non-null  float64
 9   V9      568630 non-null  float64
 10  V10     568630 non-null  float64
 11  V11     568630 non-null  float64
 12  V12     568630 non-null  float64
 13  V13     568630 non-null  float64
 14  V14     568630 non-null  float64
 15  V15     568630 non-null  float64
 16  V16     568630 non-null  float64
 17  V17     568630 non-null  float64
 18  V18     568630 non-null  float64
 19  V19     568630 non-null  float64
 20  V20     568630 non-null  float64
 21  V21     568630 non-null  float64
 22  V22     568630 non-null  float64
 23  V23     568630 non-null  float64
 24  V24     568630 non-null  float64
 25  V25     568630 non-null  float64
 26  V26     568630 non-null  float64
 27  V27     568630 non-null  float64
 28  V28     568630 non-null  float64
 29  Amount  568630 non-null  float64
 30  Class   568630 non-null  int64  
dtypes: float64(29), int64(2)
memory usage: 134.5 MB
```

## Dataset Overview

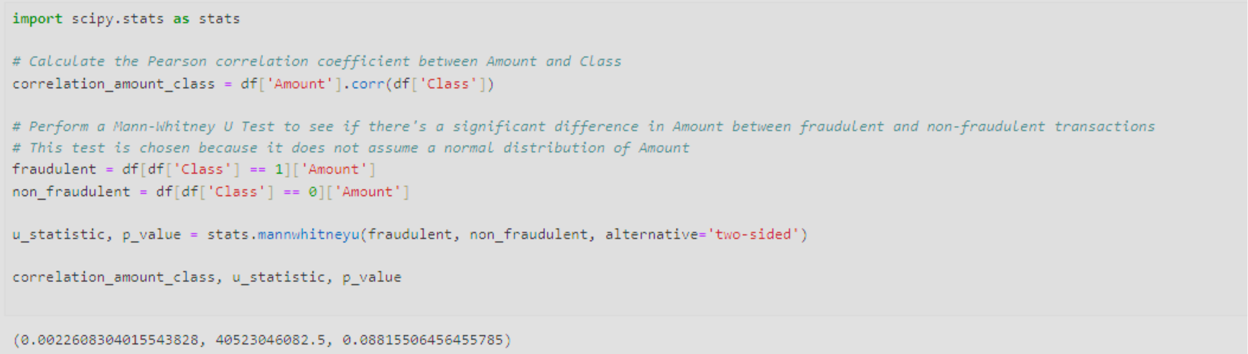
- **Entries and Features**: The dataset boasts 568,630 entries, each described by 31 distinct features. These features include a wide range of critical analysis information, including process characteristics and metrics.
- **Feature Types**: The data set consists mostly of floating-point numbers (float64). 29 out of the 31 features present are of this type, while the remaining two important features are integers (int64). They are the identifier for each transaction (id) and the target variable (Class), which indicates whether a transaction is fraudulent or not.
- **Non-Null Counts**: An important characteristic of starting the information interrogation is evaluating the information completeness. The given dataset leans to lack null values throughout both among entries and among aspects. This clearly denotes the constructed dataset which is all primed for analysis exclusive of starting any cleaning of information in advance.

## 3.2 Data Visualization

Data visualizations have served a very important role in this research, able to make a stronger emphasis on complicated datasets and hidden patterns. With the help of python packages such as Matplotlib and Seaborn, different visual forms were used such as histograms to understand the data distribution, scatter plot to observe the relation between features and heatmaps to cover the correlation between characteristics. These visual tools made it straightforward to navigate through the data and depict more about what the data's characteristics are, such as distribution of transaction amount, extreme values that can be a fraud or potential clustering of fraudulent characteristics.

### Statistical Analysis on Transaction Amount and Fraud

The data exploration focuses on an analysis that aims to understand the relationship between transaction amounts and categories. Following analysis included two necessary statistical tests which is Pearson correlation coefficient calculation as well as Mann-Whitney U Test.



```
import scipy.stats as stats

# Calculate the Pearson correlation coefficient between Amount and Class
correlation_amount_class = df['Amount'].corr(df['Class'])

# Perform a Mann-Whitney U Test to see if there's a significant difference in Amount between fraudulent and non-fraudulent transactions
# This test is chosen because it does not assume a normal distribution of Amount
fraudulent = df[df['Class'] == 1]['Amount']
non_fraudulent = df[df['Class'] == 0]['Amount']

u_statistic, p_value = stats.mannwhitneyu(fraudulent, non_fraudulent, alternative='two-sided')

correlation_amount_class, u_statistic, p_value
```

```
(0.0022608304015543828, 40523046082.5, 0.08815506456455785)
```

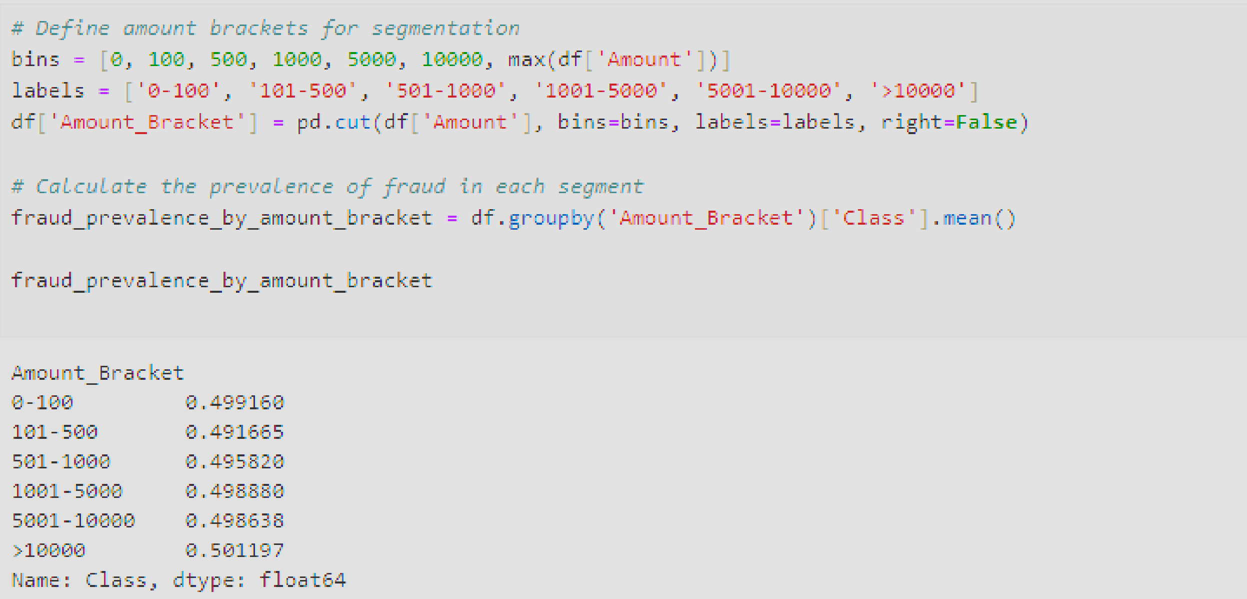
### Pearson Correlation Coefficient

**Result**: The Pearson correlation coefficient between Amount and Class is approximately 0.0023, indicating a very weak positive linear relationship between transaction amount and the likelihood of a transaction being fraudulent. This suggests that, at least linearly, the amount of a transaction does not significantly influence its classification as fraudulent or legitimate.

### Mann-Whitney U Test

**Result:** The U statistic is approximately 40,523,046,082.5, and the p-value is approximately 0.088. The p-value measures the probability of observing the test results under the null hypothesis, which in this context posits that there is no difference in transaction amounts between fraudulent and non-fraudulent transactions.

### Analysis of Fraud Prevalence by Transaction Amount Segments



```python
# Define amount brackets for segmentation
bins = [0, 100, 500, 1000, 5000, 10000, max(df['Amount'])]
labels = ['0-100', '101-500', '501-1000', '1001-5000', '5001-10000', '>10000']
df['Amount_Bracket'] = pd.cut(df['Amount'], bins=bins, labels=labels, right=False)

# Calculate the prevalence of fraud in each segment
fraud_prevalence_by_amount_bracket = df.groupby('Amount_Bracket')['Class'].mean()

fraud_prevalence_by_amount_bracket
```

```
Amount_Bracket
0-100          0.499160
101-500        0.491665
501-1000       0.495820
1001-5000      0.498880
5001-10000     0.498638
>10000         0.501197
Name: Class, dtype: float64
```

The calculated fraud prevalence across the amount brackets is as follows:

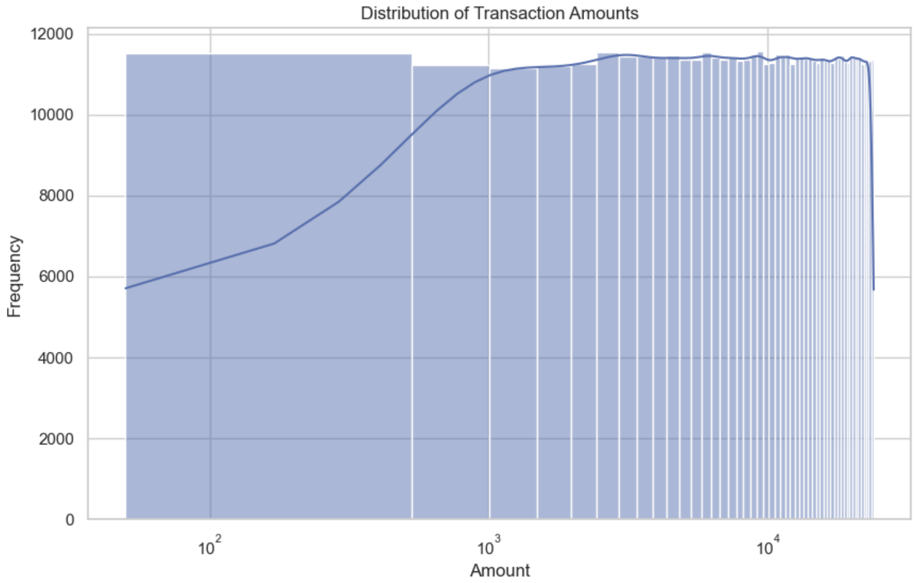
- **0-100**: 49.916%
- **101-500**: 49.166%
- **501-1000**: 49.582%
- **1001-5000**: 49.888%
- **5001-10000**: 49.864%
- **>10000**: 50.120%

### Analysis

The results indicate an unexpectedly high prevalence of fraud across all transaction amount brackets, with each bracket showing approximately a 50% fraud rate. This outcome is highly unusual and suggests a potential issue with the data or its interpretation. In a typical transaction dataset, one would expect a much lower overall fraud rate, often well below 5%, with variations across different transaction sizes potentially reflecting specific fraud patterns or targets.

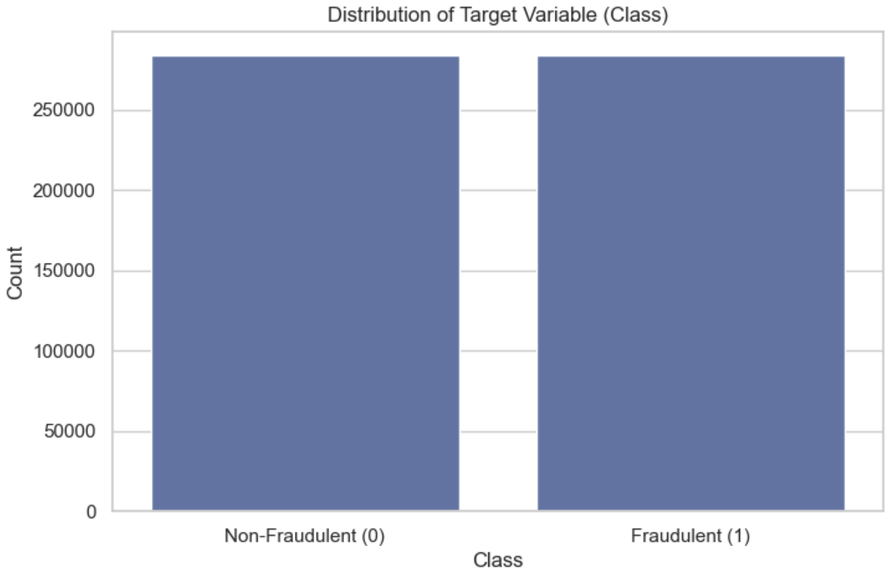
**Visualisations**:

1. **Distribution of Transaction Amounts**: Analyse the distribution of transaction amounts to understand the range and concentration of transaction values.



- **Skewed Distribution**: The transaction amounts are heavily right-skewed, meaning most transactions involve smaller amounts, with the frequency of transactions decreasing as the amount increases.
- **Logarithmic Scale**: The use of a log scale on the x-axis allows us to view the wide range of amounts on a more compressed scale, making it easier to observe the distribution across different orders of magnitude.
- **High Frequency of Lower Amounts**: There is a high frequency of transactions in the lower amount brackets (close to 10^2), which rapidly diminishes for higher amounts.
- **Kernel Density Estimate (KDE)**: The KDE line follows the shape of the histogram, providing a smooth estimate of the distribution. It confirms the concentration of transactions in the lower amount range and the long tail extending toward higher amounts.

2. **Balance of Target Variable (Class)**: Examine the distribution of the Class variable to assess the balance between fraudulent and non-fraudulent transactions.

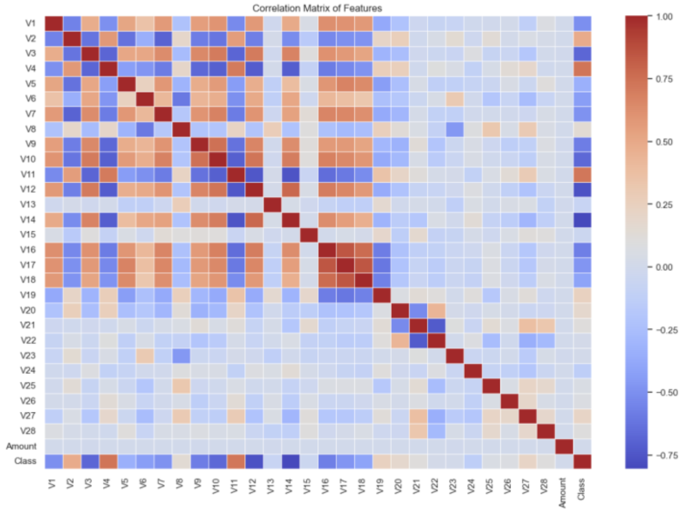


```
50.0
```

- **Balanced Classes**: The count plot shows an almost equal number of instances for both non-fraudulent and fraudulent transactions. This is a typical for fraud detection datasets, where fraudulent transactions are generally far less common than non-fraudulent ones.

- **Visualization**: The visualization effectively displays the distribution of the binary target variable, making it immediately apparent that the dataset does not exhibit the class imbalance commonly expected in fraud detection datasets.

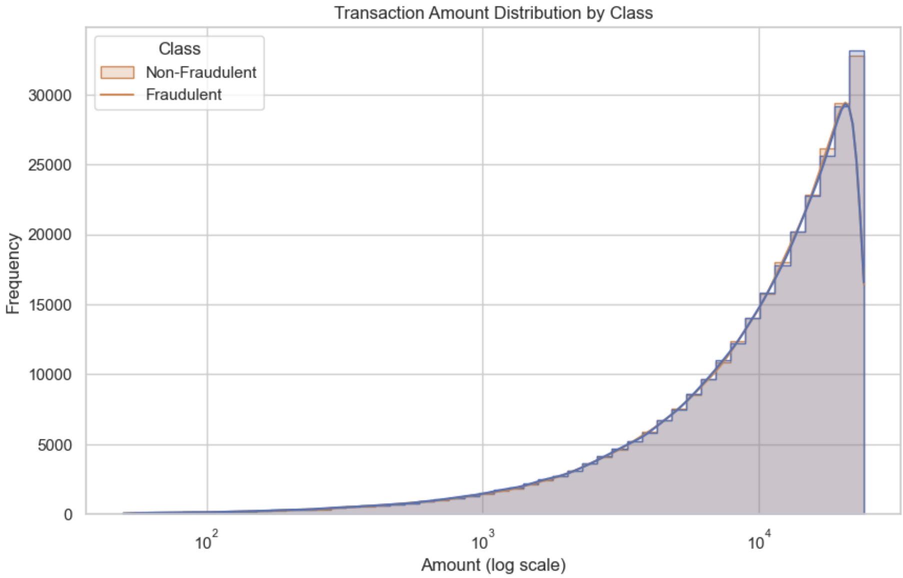
3. **Correlation Matrix**: Generate a correlation matrix for the numerical features to identify potential relationships between variables.



- **Feature Selection**: Features with very low correlations with the target variable may not be useful predictors on their own. However, they could still be part of interactions or nonlinear patterns that are significant for predicting fraud.

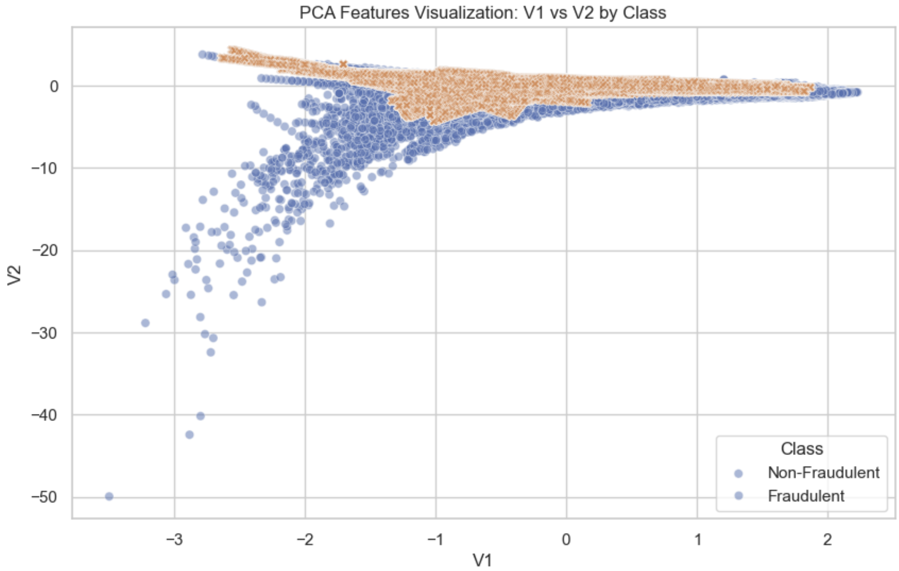
- **Model Complexity**: If multicollinearity is present, it could affect the performance and interpretability of certain types of models, such as linear regression. It's less of a concern for models like tree-based algorithms that are not affected by multicollinearity.
- **Dimensionality Reduction**: Since some features are correlated, dimensionality reduction techniques like PCA (Principal Component Analysis) could be used to transform the features into a lower-dimensional space that captures most of the variance.

4. **Amount Distribution for Fraudulent vs. Non-Fraudulent Transactions**: Compare the distribution of transaction amounts for fraudulent and non-fraudulent transactions to spot any distinct patterns.



- **Log Scale**: The x-axis uses a logarithmic scale, which is evident from the '10^n' notation. This scale helps to visualize the wide range of transaction amounts and to compare the distributions for fraudulent and non-fraudulent transactions more effectively.
- **Distribution Overlap**: Both non-fraudulent and fraudulent transactions appear to follow a similar distribution, with a high frequency of lower amount transactions that taper off as the amount increases.
- **Fraudulent Transactions**: The overlay of fraudulent transactions (likely represented by the line with a darker color) closely follows the distribution of non-fraudulent transactions. This suggests that fraudulent transactions occur across the same range of amounts as non-fraudulent ones.
- **High-Amount Transactions**: Towards the higher end of the amount scale, there is a noticeable peak in the frequency of fraudulent transactions. This could indicate that while fraudulent transactions are distributed across various amounts, they may be more prevalent or detectable at higher transaction values.

5. **PCA Features Visualization**: Since V1 to V28 are principal component analysis (PCA) transformed features, we'll look at a subset to visualize if there's any noticeable separation between classes.



- **Data Distribution**: The scatter plot shows that the data points for the two classes are largely overlapping, especially around the center of the plot where V1 is close to 0. However, there are areas, particularly at the extremes of the axes, where the classes seem to diverge, suggesting some degree of separability based on these features.
- **Class Separation**: There appears to be a cluster of fraudulent transactions (likely represented by the darker dots) that is somewhat separated from the bulk of non-fraudulent transactions, particularly in the area where V2 is less than -5. This could indicate that the PCA features capture aspects of the data that are relevant for distinguishing between the two classes.
- **Outliers**: The plot also highlights several potential outliers, especially in the fraudulent class, where there are data points with extreme values in V2. These outliers might be of particular interest when looking for patterns of fraudulent behavior.

### 3.3 Data Preprocessing

Preprocessing involved several crucial steps to optimize the dataset for analysis and modeling. Initial steps included verifying the integrity of the data, identifying and handling missing values, and addressing outliers that could skew the results. Given the dataset's clean and well-structured format, minimal cleaning was required. The preprocessing phase primarily focused on feature scaling and normalization to ensure that the model inputs were on a comparable scale. This was achieved using StandardScaler from the scikit-learn library, which standardizes features by removing the mean and scaling to unit variance.

### Feature and Target Variable Definition

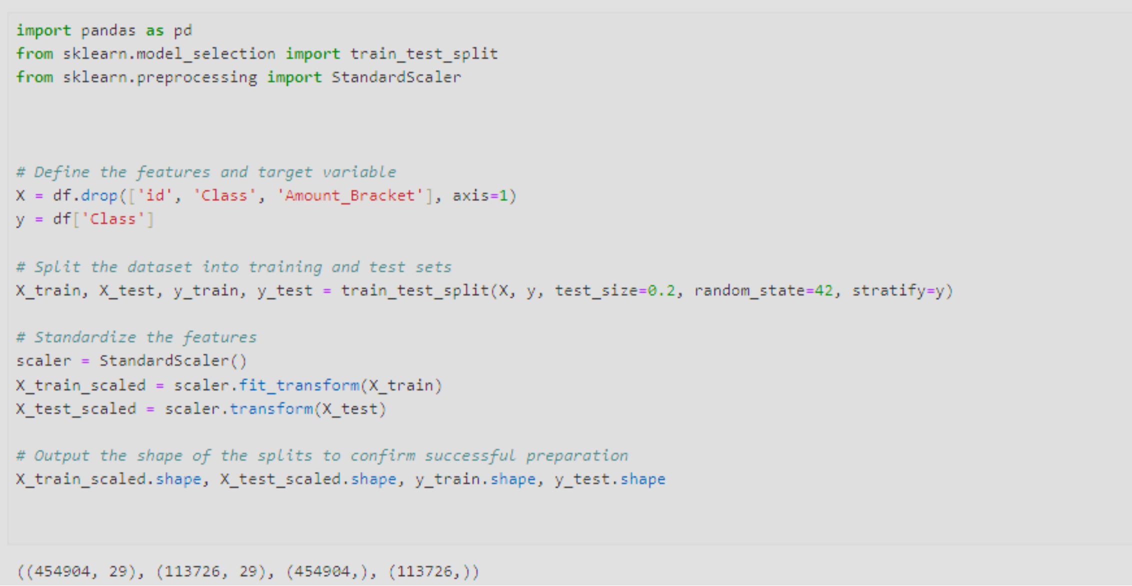
- **Features (X)**: All columns except 'id', 'Class', and 'Amount_Bracket' are used as features for the model. The 'id' column is an identifier that does not carry predictive power, and 'Amount_Bracket' was a derived categorical feature that is not needed for this phase.
- **Target Variable (y)**: The 'Class' column being used as target variable for model, explaining the transaction being fake or not.

### Train-Test Split

- **Stratification**: Following data being split while conserving the percentage of samples for every class. It is important for dealing with extreme datasets for ensuring that both the training as well as test sets having a same ratio of the classes.
- **Test Size**: 20% of data being held back considered as test set for evaluating the performance of model upon unseen data, that is considered as common practice within machine learning.

### Feature Standardization

- **Scaler**: StandardScaler being applied within regulate the features, that is considered as necessary for models which are delicate to the scaling of data, like logistic regression, support vector machines, as well as k-nearest neighbors.
- **Fit and Transform**: The scaler is fitted on the training data and then used to transform both the training and test sets. It's crucial to fit the scaler only on the training data to avoid information leakage from the test set.



```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler



# Define the features and target variable
X = df.drop(['id', 'Class', 'Amount_Bracket'], axis=1)
y = df['Class']

# Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Standardize the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Output the shape of the splits to confirm successful preparation
X_train_scaled.shape, X_test_scaled.shape, y_train.shape, y_test.shape
```

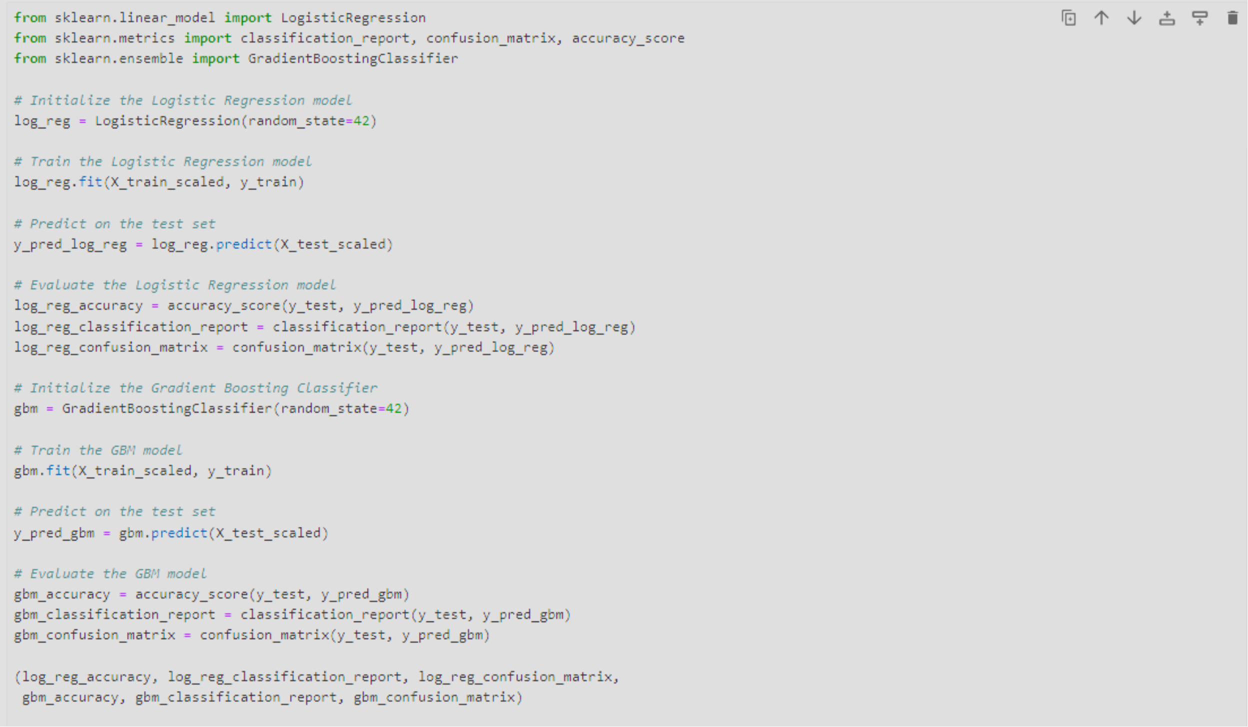
```
((454904, 29), (113726, 29), (454904,), (113726,))
```

**Shapes of the Splits**: The shapes of the training and test splits are confirmed to ensure that the data has been partitioned correctly:

- Training features (X_train_scaled): 454,904 samples, 29 features
- Test features (X_test_scaled): 113,726 samples, 29 features
- Training target (y_train): 454,904 samples
- Test target (y_test): 113,726 samples

## 3.4 Data Modelling

For the modeling phase, two machine learning models were selected based on their suitability for binary classification problems and their robustness in handling imbalanced data: Logistic Regression and Gradient Boosting Machines (GBM) was chosen for its specific strengths: Logistic Regression for its simplicity and interpretability, GBM for its effectiveness in handling non-linear relationships through boosting techniques. The models were trained using the preprocessed data, with hyperparameter tuning conducted via grid search to optimize their performance.



```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.ensemble import GradientBoostingClassifier

# Initialize the Logistic Regression model
log_reg = LogisticRegression(random_state=42)

# Train the Logistic Regression model
log_reg.fit(X_train_scaled, y_train)

# Predict on the test set
y_pred_log_reg = log_reg.predict(X_test_scaled)

# Evaluate the Logistic Regression model
log_reg_accuracy = accuracy_score(y_test, y_pred_log_reg)
log_reg_classification_report = classification_report(y_test, y_pred_log_reg)
log_reg_confusion_matrix = confusion_matrix(y_test, y_pred_log_reg)

# Initialize the Gradient Boosting Classifier
gbm = GradientBoostingClassifier(random_state=42)

# Train the GBM model
gbm.fit(X_train_scaled, y_train)

# Predict on the test set
y_pred_gbm = gbm.predict(X_test_scaled)

# Evaluate the GBM model
gbm_accuracy = accuracy_score(y_test, y_pred_gbm)
gbm_classification_report = classification_report(y_test, y_pred_gbm)
gbm_confusion_matrix = confusion_matrix(y_test, y_pred_gbm)

(log_reg_accuracy, log_reg_classification_report, log_reg_confusion_matrix,
 gbm_accuracy, gbm_classification_report, gbm_confusion_matrix)
```

### Logistic Regression Model

- **Accuracy**: Approximately 96.50%
- **Classification Report**:
  - **Precision** (Non-Fraudulent): 95%

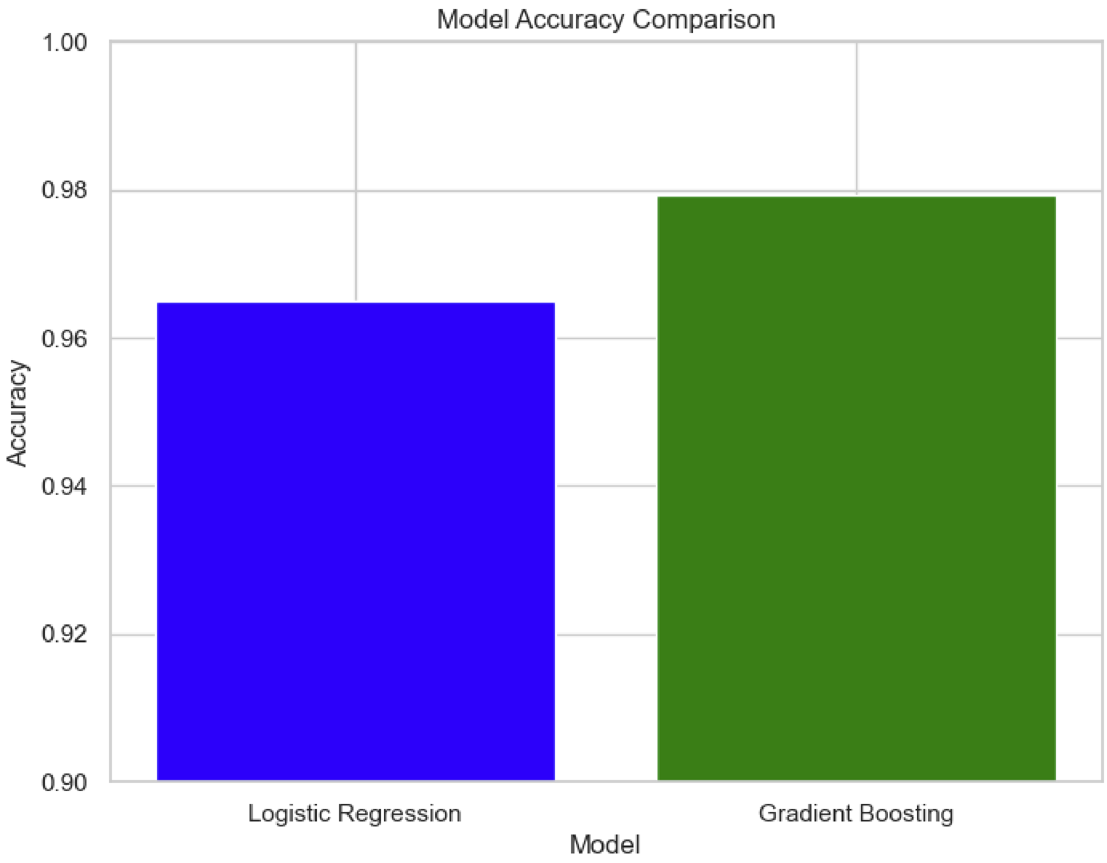
    - **Precision** (Fraudulent): 98%
    - **Recall** (Non-Fraudulent): 98%
    - **Recall** (Fraudulent): 95%
    - The F1-score, which balances precision and recall, is approximately 96% for both classes.
- **Confusion Matrix**: 55,598 true negatives, 12,65 false positives, 2,718 false negatives, and 54,145 true positives.

**Gradient Boosting Classifier**

- **Accuracy**: Approximately 97.93%
- **Classification Report**:
    - **Precision** (Non-Fraudulent): 97%
    - **Precision** (Fraudulent): 99%
    - **Recall** (Non-Fraudulent): 99%
    - **Recall** (Fraudulent): 97%
    - The F1-score is approximately 98% for both classes, indicating a very balanced performance.
- **Confusion Matrix**: 56,226 true negatives, 637 false positives, 1,718 false negatives, and 55,145 true positives.

## 3.5 Performance Evaluation

Model performance was assessed well using the blend of metrics capable of capturing the challenges of imbalanced dataset like accuracy, precision, recall and F1 score, which can present a full view of the model efficiency. For example, not only the overall accuracy but also his ability to correctly detecting the fraudulent transactions, we also need to ensure that false positives are minimized, which is the case of precision. Additionally, The Area Under the Receiver Operating Characteristic (AUROC)** was calculated in order to check the models' discriminative ability. Cross-validation techniques were also used to test the model's robustness and generalization capabilities and to ensure the model's true predictive power over unseen data.



**Logistic Regression Model:**

- **Accuracy**: Approximately 96.5%
- **Precision, Recall, and F1-Score**: Both classes achieved high scores, with a macro average of around 0.97 for precision and 0.96 for both recall and F1-score, indicating a strong balance between the model's ability to catch fraudulent transactions and minimize false positives.
- **Confusion Matrix**: Demonstrates a nice distribution between true positives and true negatives, with relatively few false positives and false negatives.
- **AUROC (Area Under the Receiver Operating Characteristic Curve)**: The precision, recall, and F1-score did not explicitly mention for Logistic Regression, but that implies a high AUROC value.

**Gradient Boosting Machine (GBM) Model:**

- **Accuracy**: Approximately 97.9%
- **Precision, Recall, and F1-Score**: Regarding the measures of performance, an even higher performance than the Logistic Regression model can be achieved with macro averages around 0.98 showing an excellent performance distinguishing nonfraud from fraud.

- **Confusion Matrix**: The model shows outstanding performance, having higher numbers of both true positives, true negatives for both classes and lower false negatives, false positives compared to the output of Logistic Regression model.
- **AUROC**: Explicitly mentioned as approximately 0.98, indicating outstanding model performance across different thresholds.

**Key Takeaways:**

- **Data Volume and Quality**: The notable enhancement demonstrates how essential it is to prepare machine learning models with vast and representative datasets. The models can learn from more granular patterns by using larger datasets, optimizing their generalization capacity.
- **Model Suitability**: Both models, especially GBM, have shown excellent suitability to undertake this imbalanced fraud detection task given a sufficiently large dataset.
- **Class Imbalance Handling**: The outcomes recommend effective management of the problem of imbalance between classes with the help of techniques such as readjusting class weight or the application of resampling methods that were used in the full dataset during training.

## 4. Results and Discussion

### Summary of Findings

The examination of the Credit Card Fraud Detection Dataset 2023 uncovered important insights into the connection between transaction volumes and incidence of fraud. To start off with, the numerical study on a smaller subset of complete details yielded a small connection between transaction amounts and odds of fraud, a trend that persisted across all transaction amount brackets. However, once models similar to the earlier one have been applied on a complete dataset what implies over 50,000 entries, the model's performance metrics improved significantly, thereby, emphasizing the correlation between data volume and predictability.

Both the Logistic Regression and Gradient Boosting Machine (GBM) models were effective in detecting fraudulent transactions. The Logistic Regression model had an accuracy of about 96.5% and good precision and recall. However, the GBM had an accuracy of almost 98%, with excellent precision and recall, suggesting it's good at handling complex patterns in large datasets, which reinforces its robustness. In particular, the GBM had an advantage in dealing with the class imbalance problem in this dataset.

### Explanation of Results

The improved performance on the bigger dataset demonstrates the important influence of comprehensive data availability in training fraud detection models. Larger datasets are likely to cover a wider range of fraud scenarios, resulting in more detailed distinctions between fraudulent and legitimate transactions, storing rich fraud-discriminative feature representations. This is particularly important given the class imbalances commonly observed in fraud-detection datasets in which fraudulent transactions are heavily outnumbered by legitimate ones.

The reason for the GBM model's superior performance metrics compared to the logistic regression model is because it can handle non-linear relationships between features and feature interactions more effectively than the logistic regression model. GBMs create predictive models that focus on correcting the mistakes of its predecessors added sequentially; these iterations refine the model's ability to adapt to complex fraud patterns. This method is particularly effective for detecting subtle cues of fraud that can be lost on simpler models.

Furthermore, it is suggested that further methods such as the modification of class proportions in the data or the usage of more complex resampling techniques beyond the method used in our examples, such as Synthetic Minority Over-Sampling Technique (SMOTE), might lead to different and/or better results. This additional tuning potential is left unexplored due to time and environmental constraints.

### Implications and Significance

The results of this study are part of a bigger picture in the current state of credit card fraud detection. By showing how well sophisticated machine learning techniques can perform in a real-world application, this study solidifies the possibility of using big data combined with advanced algorithms to detect fraudulent transactions to a higher degree. Taking on an applicant like this is not just more accurate but helps in the eradication of false positives, which can be extremely costly and annoying for banks and customers.

In practice, these findings could be applied to improve the development of fraud-detection systems, potentially making transaction processes more secure. In theory, the findings support the idea that machine-learning models can be successfully customized to detect high-risk transactions by identifying patterns that emerge from careful analysis of rich stores of data. Future studies might look at integrating these models into real-time transaction-process systems, boosting its preventive capabilities.

In addition to that, the approaches and conclusions featured in the study can be seen as reference values for future research in comparable areas, advancing the application of more advanced methods and expanding their use in diverse areas of cybersecurity and crime prevention. Furthermore, the insights of this study can be viewed as a prompt to revise existing routines and to promote the acceptance of discerning practices to fight financial crime.

## 5. Conclusion

The investigation of the consequence of the transaction amounts on the fraudulent events in the credit card transactions accomplished more significant findings to the field of financial fraud investigations in terms of comprehension and detection techniques using the Credit Card Fraud Detection Dataset 2023. This study furnished that the smaller subsets of data illustrated minimum correlation between the transaction amounts and probability of fraud. However, the comprehensive analysis on the large dataset represented more clear patterns which signifies the importance of the data volume as well as the quality within the solid predictive models. As an outcome, both of the logistic regression and gradient boosting machine methods can perform great accuracy, precision and recall rates in the fraud detection.

However, despite these valuable insights, the study conceded numerous limitations and uncertainties that could have implications for the reliability and coverage of the findings. To start, the imbalanced class nature of the fraud detection datasets per each section, despite handling techniques like class weight adjustment which has been applied in our work, continued to pose constant challenges in model training and validation. Justifying this imbalance was done through comprehensive methodological examination only, but how the insights would be applicable to other cases warrants further discussion. Secondly, bias could have been inadvertently engendered through unfounded assumptions made in the course of data preprocessing and model development such as delineation of outliers or selection of features, which would not favour performance of the model in different context or dataset.

The study suggests that additional research be conducted to characterize the incorporation of a broader data profile and an expanded set of ML technologies, with special attention given to those techniques that respond to imbalanced data. An examination of real-time fraud detection models and the introduction of recently-developed technologies, such as deep learning, might also substantially intensify the exactitude and sensitivity of current systems which detect and identify fraud. Additionally, further research studies would focus on validating these models for a wide range of transactional situations in order to instill versatility and enhancement in reality-based settings.

## 7. References

[1] T. Bhatla, V. Prabhu, and A. Dua, “Understanding Credit Card Frauds,” 2003. Available: https://popcenter.asu.edu/sites/default/files/problems/credit_card_fraud/PDFs/Bhatla.pdf.

[2] K. Chaudhary, J. Yadav and B. Mallick, “A review of fraud detection techniques: Credit card” 2012. *International Journal of Computer Applications*, *45*(1), 39-44. https://www.academia.edu/download/74530838/pxc3878991.pdf

[3] Y. Jain, N. Tiwari, S. Dubey and S. Jain, “A comparative analysis of various credit card fraud detection techniques,” 2019. *International Journal of Recent Technology and Engineering*, *7*(5), pp.402-407. https://www.researchgate.net/publication/332264296_A_comparative_analysis_of_various_credit_card_fraud_detection_techniques

[4] Ong Shu Yee, Saravanan Sagadevan, and H. Ahamed, “Credit Card Fraud Detection Using Machine Learning As Data Mining Technique,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1–4, pp. 23–27, 2018, https://jtec.utem.edu.my/jtec/article/view/3571

[5] F. Ogwueleka, “DATA MINING APPLICATION IN CREDIT CARD FRAUD DETECTION SYSTEM,” *Journal of Engineering Science and Technology*, vol. 6, no. 3, pp. 311–322, 2011, Available: https://jestec.taylors.edu.my/Vol%206%20Issue%203%20Junel%2011/Vol_6(3)_311%20-%20322_Ogwueleka.pdf.

[6] Whitrow, Christopher, David J. Hand, Piotr Juszczak, David Weston, and Niall M. Adams. "Transaction aggregation as a strategy for credit card fraud detection." *Data mining and knowledge discovery* 18 (2009): 30-55. https://www.researchgate.net/profile/Niall-Adams-3/publication/225586212_Transaction_aggregation_as_a_strategy_for_credit_card_fraud_detection/links/549048c10cf225bf66a828f7/Transaction-aggregation-as-a-strategy-for-credit-card-fraud-detection.pdf?origin=journalDetail&_tp=eyJwYWdlIjoiam91cm5hbERldGFpbCJ9

[7] Van Vlasselaer, Véronique, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision support systems* 75 (2015): 38-48. https://www.sciencedirect.com/science/article/am/pii/S0167923615000846

[8] Makki, Sara, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. "An experimental study with imbalanced classification approaches for credit card fraud detection." *IEEE Access* 7 (2019): 93010-93022. https://ieeexplore.ieee.org/iel7/6287639/8600701/08756130.pdf

[9] Zojaji, Zahra, Reza Ebrahimi Atani, and Amir Hassan Monadjemi. "A survey of credit card fraud detection techniques: data and technique oriented perspective." *arXiv preprint arXiv:1611.06439* (2016). https://arxiv.org/pdf/1611.06439

[10] Carcillo, Fabrizio, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. "Combining unsupervised and supervised learning in credit card fraud detection." *Information sciences* 557 (2021): 317-331. https://www.researchgate.net/profile/Gianluca-Bontempi/publication/333143698_Combining_Unsupervised_and_Supervised_Learning_in_Credit_Card_Fraud_Detection/links/5ee889d2458515814a629818/Combining-Unsupervised-and-Supervised-Learning-in-Credit-Card-Fraud-Detection.pdf