**TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC QUỐC GIA HÀ NỘI**

**KHOA KẾ TOÁN KIỂM TOÁN**

**--------------------------**

**Logo

Description automatically generated**

**BUSINESS ANALYST IN ACCOUNTING**

***MID TERM:***

# Credit Card Fraud Detection and Building Model

|  |  |  |
| --- | --- | --- |
| **Sinh viên thực hiện** | **:** | **NGUYỄN THỊ THU HUYỀN (22050747)** |
| **Khoa** | **:** | **Kế toán – Kiểm toán** |
| **Lớp** | **:** | **QH 2022E – Kế toán 1** |

# *Hà Nội – 04/2025*

[**Table of Contents** 1](#_Toc195792772)

[**1 Introduction** 1](#_Toc195792773)

[1.1 Research Objective 3](#_Toc195792774)

[1.2 The Importance of Credit Card Fraud Detection 3](#_Toc195792775)

[**2 Data Selection and Description** 1](#_Toc195792773)

[2.1 Dataset Introduction 4](#_Toc195792776)

[2.2 Data Characteristics 4](#_Toc195792776)

[2.3 Reason for Choosing the Dataset 4](#_Toc195792777)

[**3 Research Questions and Analytical Objectives** 12](#_Toc195792778)

[3.1 Research Questions 4](#_Toc195792776)

[3.2 Analytical Objectives 4](#_Toc195792776)

[**4 Data Analysis, Model Development, and Evaluation** 12](#_Toc195792778)

[4.1 Data Preprocessing 4](#_Toc195792776)

[4.2 Exploratory Data Analysis (EDA) 4](#_Toc195792776)

[4.3 Building the Random Forest Model 4](#_Toc195792776)

[4.4 Model Results 4](#_Toc195792776)

[**5 Ethical Considerations** 13](#_Toc195792779)

[5.1 Customer Data Privacy and Anonymization 13](#_Toc195792780)

[5.2 Transparency in Model Usage to Avoid Bias or Discrimination 13](#_Toc195792781)

[**6 Conclusion and Future Directions** 21](#_Toc195792784)

[6.1 Summary of Achievements 13](#_Toc195792780)

[6.2 Limitations and Suggestions for Improvement 13](#_Toc195792781)

**Appendix**  [1](#_Toc195792772)

### 

**1. Introduction**  
**1.1. Research Objective**  
This research aims to **analyze** credit card transaction data and **build a machine learning model** to detect fraudulent transactions. By exploring the characteristics of the data and applying modern algorithms, the study aims to enhance the effectiveness and accuracy of anomaly detection.

**1.2. Importance of Credit Card Fraud Detection**  
Credit card fraud is a **serious issue** in the financial and banking sector, causing significant losses for both individuals and organizations. Early detection of fraudulent activities not only helps to improve the security of payment systems but also supports the monitoring and auditing of financial activities. This can play a key role in accounting and auditing processes to ensure transparency and regulatory compliance.

**2. Data Selection and Description**   
**2.1. Dataset Introduction: Credit Card Fraud Detection Dataset (Source: Kaggle)**  
The dataset contains transactions made by credit cards in September 2013 by European cardholders.  
This dataset presents transactions that occurred over two days, with 492 frauds out of 284,807 transactions. **The dataset is highly imbalanced**, with the positive class (frauds) accounting for 0.172% of all transactions.  
It contains only numerical input variables, which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features or more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, and the only features that have not been transformed with PCA are 'Time' and 'Amount'. The 'Time' feature contains the seconds elapsed between each transaction and the first transaction in the dataset. The 'Amount' feature represents the transaction amount, which can be used for example-dependent cost-sensitive learning. The 'Class' feature is the response variable, taking a value of 1 for fraud and 0 otherwise.  
Given the class imbalance ratio, we recommend measuring accuracy using the Area Under the Precision-Recall Curve (AUPRC), as confusion matrix accuracy is not meaningful for imbalanced classification.

**2.2. Data Characteristics**

* **Number of transactions**: The dataset contains a total of 284,807 transactions, out of which approximately 492 are fraudulent transactions. This presents a significant class imbalance, with fraudulent transactions constituting only about 0.172% of all transactions. This imbalance makes the task of fraud detection particularly challenging, as traditional machine learning algorithms may struggle to identify the rare fraudulent cases without appropriate techniques to handle the class imbalance.
* **Feature Variables**:
  + **V1–V28**: These are the principal components derived from a PCA (Principal Component Analysis) transformation applied to the original dataset. Due to confidentiality issues, the original features and their meanings are not provided. However, V1 through V28 represent abstract, anonymized features that capture the underlying patterns in the data. These features are all numerical and are used as input variables for the machine learning model.
  + **Amount**: This feature represents the monetary value of each transaction. The 'Amount' variable can be used for example-dependent, cost-sensitive learning, allowing the model to account for larger transactions, which may carry higher risks of fraud.
  + **Time**: The 'Time' feature captures the amount of time (in seconds) that has passed since the first transaction in the dataset. This variable is useful for detecting patterns in the sequence of transactions, as fraudulent activities may exhibit different temporal behaviors compared to legitimate transactions.
  + **Class**: This is the target variable that indicates whether a transaction is fraudulent or not. A value of 1 denotes a fraudulent transaction, while a value of 0 indicates a legitimate one.
* **Dataset Size**: The dataset consists of 284,807 rows (transactions) and 31 columns (including the feature variables and the target variable). This large volume of data allows for robust model training, though the imbalanced nature of the data requires special attention in terms of preprocessing and evaluation.

**2.3. Reasons for Choosing the Dataset**  
The **Credit Card Fraud Detection** dataset was selected for this research for several reasons related to its practical relevance, alignment with machine learning objectives, and application value in anomaly detection:

* **High Class Imbalance**: This is a prominent and highly realistic characteristic of fraud detection problems. In this dataset, the number of legitimate transactions (class = 0) overwhelmingly exceeds the number of fraudulent transactions (class = 1), with frauds accounting for only about 0.17%. Handling and building models on such an imbalanced dataset is a typical challenge in machine learning, helping students and researchers gain a better understanding of how to tackle similar real-world scenarios.
* **Suitability for Anomaly Detection Methods**: Since fraud cases are rare and exhibit different behavioral characteristics, this dataset is highly suitable for experimenting with anomaly detection models, such as Isolation Forest, One-Class SVM, Autoencoder, and more. These models are designed to identify unusual or rare events, making them ideal for fraud detection.
* **Real-World and Preprocessed Data**: The dataset, provided by researchers, originates from actual credit card transactions in Europe in 2013. It has been preprocessed using PCA (Principal Component Analysis) to ensure the confidentiality and anonymization of personal information, while still retaining key features important for analysis.
* **Reasonable Distribution of Time and Transaction Amount**: The two features, "Time" and "Amount", are not encoded, allowing for visualization and analysis of user behavior over time or according to transaction value. This is an important factor in identifying unusual behaviors and fraud patterns based on these attributes.
* **High Applicability**: The techniques and insights derived from analyzing this dataset can be extended and applied in real-world scenarios, not only for financial and banking surveillance systems but also to support accounting and auditing departments in detecting irregularities, safeguarding assets, and ensuring transparency in transactions.

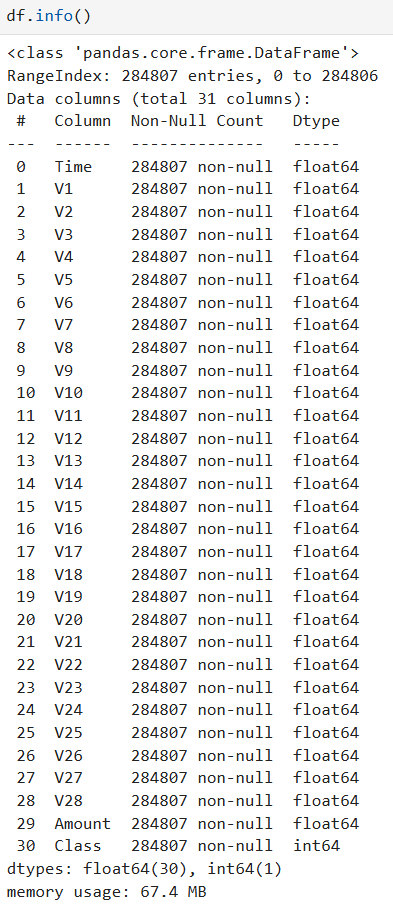
**3. Research Questions and Analysis Objectives**   
**3.1. Research Questions**  
To address the problem of fraud detection in credit card transactions—an issue that is highly practical and challenging—the research team poses the following main questions:

* How can fraudulent transactions be detected accurately, given that the fraud rate in the data is very small?
* Can the features encoded in the data help identify anomalous behaviors?
* Which machine learning model is most suitable for a problem with such a high imbalance like this?
* Can deep learning models, such as Autoencoder, outperform traditional models in fraud detection?

**3.2. Analysis Objectives**  
Based on the above research questions, the team identifies the following analysis objectives:

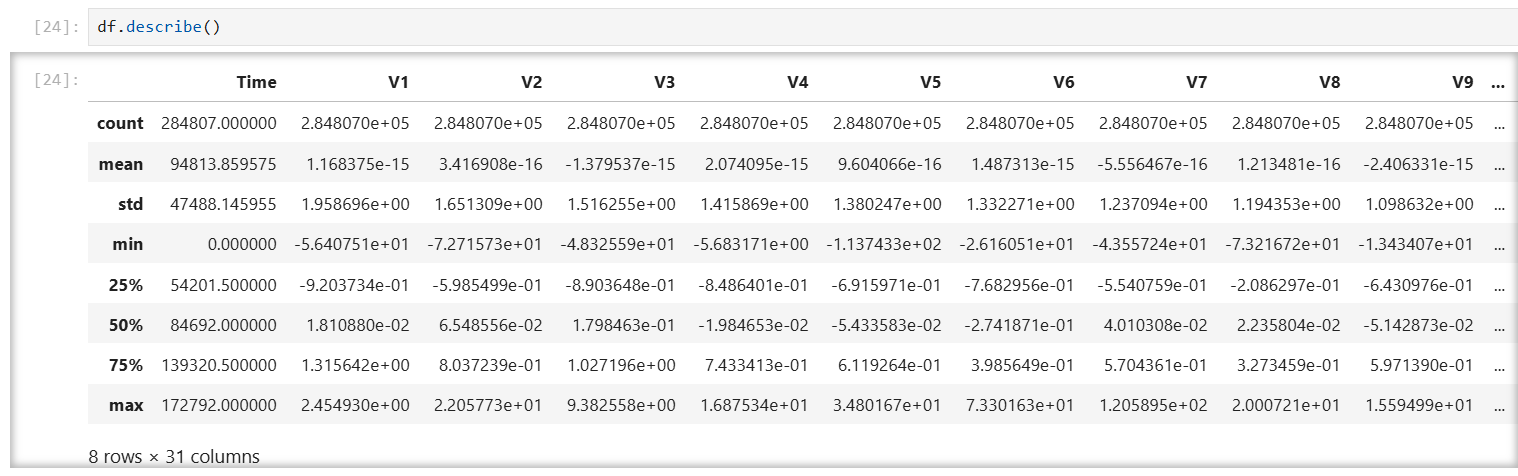
* **Analyze the transaction data**:
  + Explore the distribution characteristics of legitimate and fraudulent transactions.
  + Visualize the relationship between time, transaction amount, and the likelihood of fraud.
  + Examine the class imbalance in the data and its impact on model evaluation metrics.
* **Build a fraud detection model**:
  + Apply unsupervised learning techniques, specifically Autoencoder, to learn the patterns of legitimate transactions and detect anomalous transactions based on reconstruction error.
  + Compare the performance of Autoencoder with traditional models such as Logistic Regression, Random Forest, or Isolation Forest.
* **Evaluate model effectiveness**:
  + Use evaluation metrics suitable for imbalanced problems, such as Precision, Recall, and F1-score, rather than relying solely on Accuracy, which can be misleading in anomaly detection tasks.

**4. Analysis and Evaluation of Model Results**   
**4.1. Data Preprocessing**

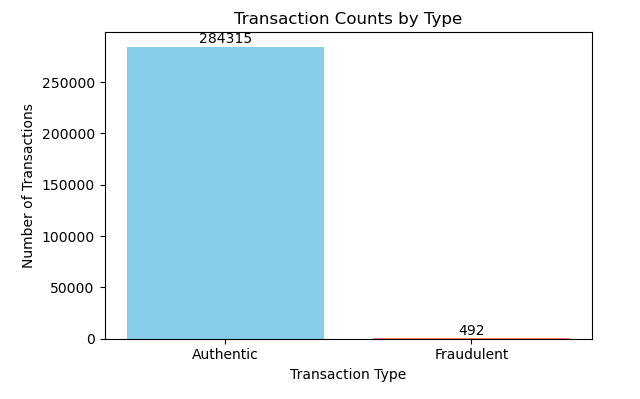


The dataset does not contain any missing values, and the data types of the columns are appropriate for analysis.

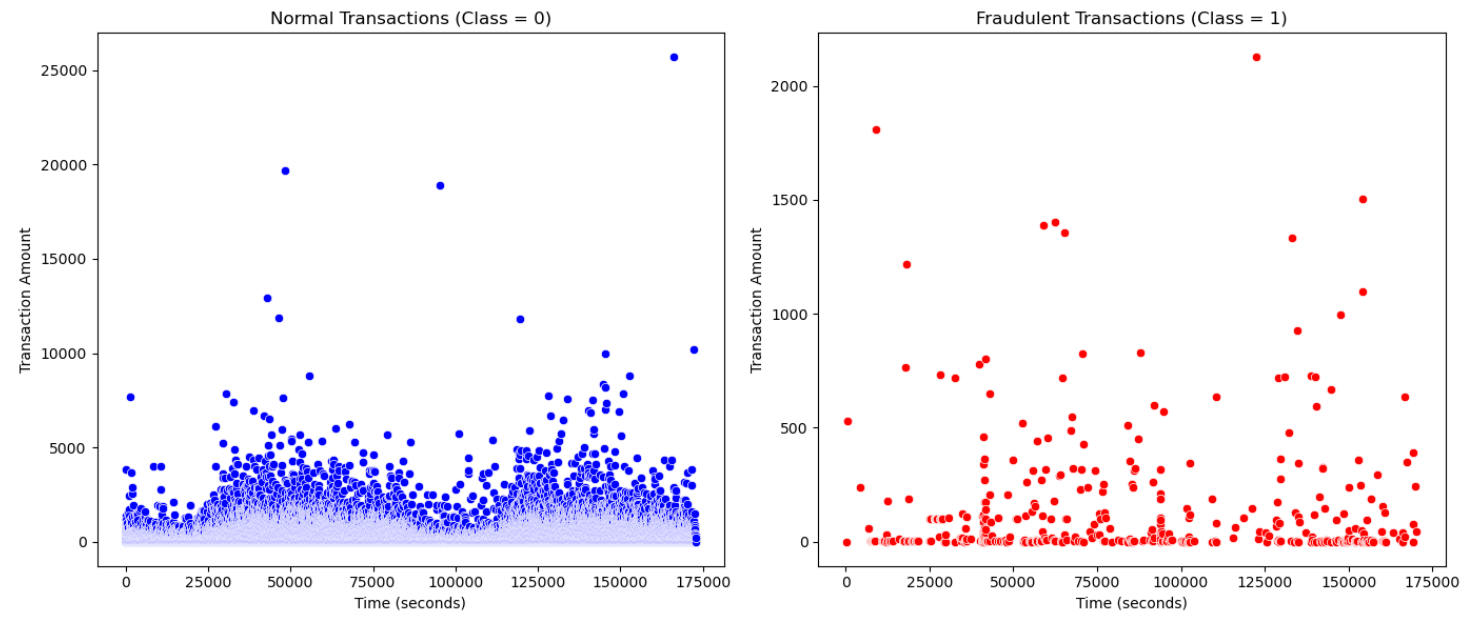
* An overview of the dataset indicates that it has been thoroughly cleaned and preprocessed, making it suitable for further analysis and exploration. The cleaning process ensures that the data is free from inconsistencies, such as missing or erroneous values, and that all features are in the correct format for effective modeling. This clean and well-prepared dataset provides a solid foundation for performing in-depth data analysis and implementing machine learning models.

**4.2. Exploratory Data Analysis – EDA**  


Looking at the **Time** feature, we can confirm that the dataset contains 284,807 transactions, spanning across a period of 2 consecutive days, which is equivalent to 172,792 seconds. This temporal aspect is important as it allows us to analyze transaction patterns within a specific time frame, providing insights into whether fraud is more likely to occur at certain times of the day or during particular intervals. By examining the distribution of transaction times, we can identify any unusual activity or spikes in transaction volume that may correspond with fraudulent behavior. This analysis is a critical part of understanding how temporal variables might influence the detection of anomalies in the dataset.



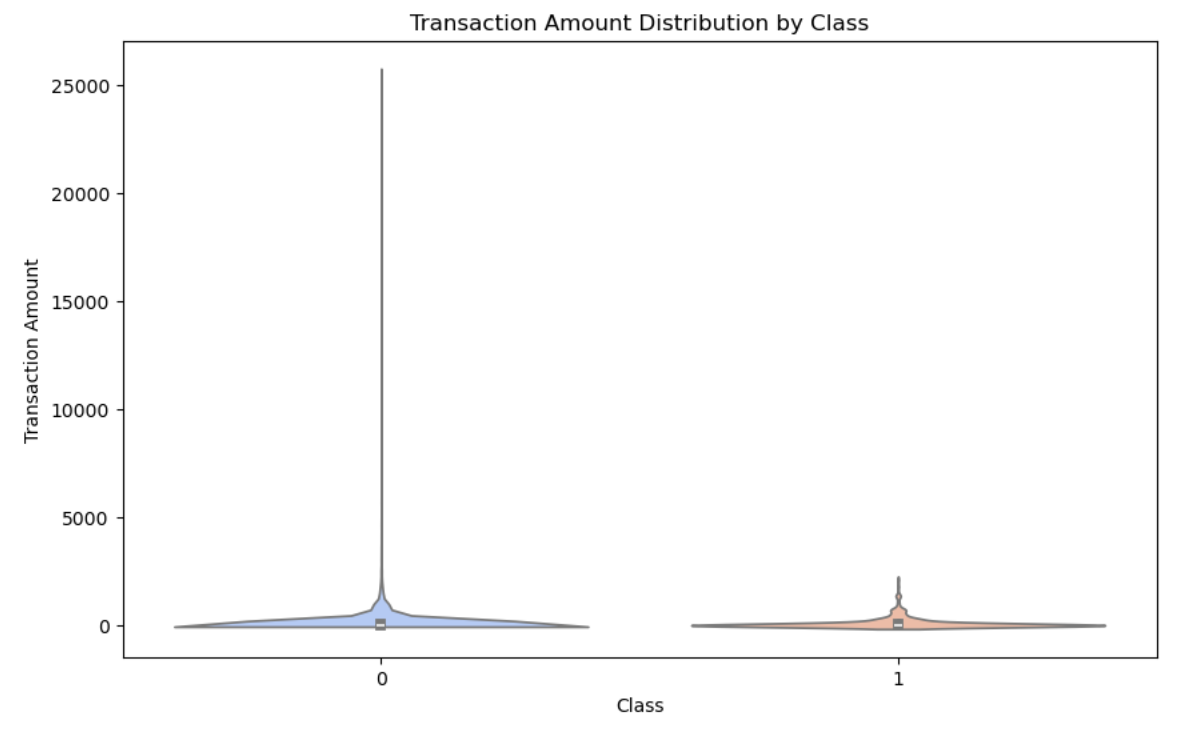
Only 492 (or 0.172%) of the transactions are fraudulent. This indicates that the dataset is highly imbalanced with respect to the target variable, **Class**. The overwhelming majority of transactions are legitimate (Class = 0), while fraudulent transactions (Class = 1) are extremely rare. This class imbalance presents a challenge for machine learning models, as they may struggle to correctly classify the minority class (fraudulent transactions) without using special techniques to address the imbalance, such as resampling, cost-sensitive learning, or appropriate evaluation metrics.



**Legitimate transactions (Class = 0)** are abundant, evenly distributed over time, and show a wide range of transaction amounts, including many large transactions (above 5,000 and even over 20,000).

**Fraudulent transactions (Class = 1)** occur less frequently and are mostly small transactions (under 1,000). However, these fraudulent transactions appear sporadically throughout the entire time period, with no concentration in any specific phase.

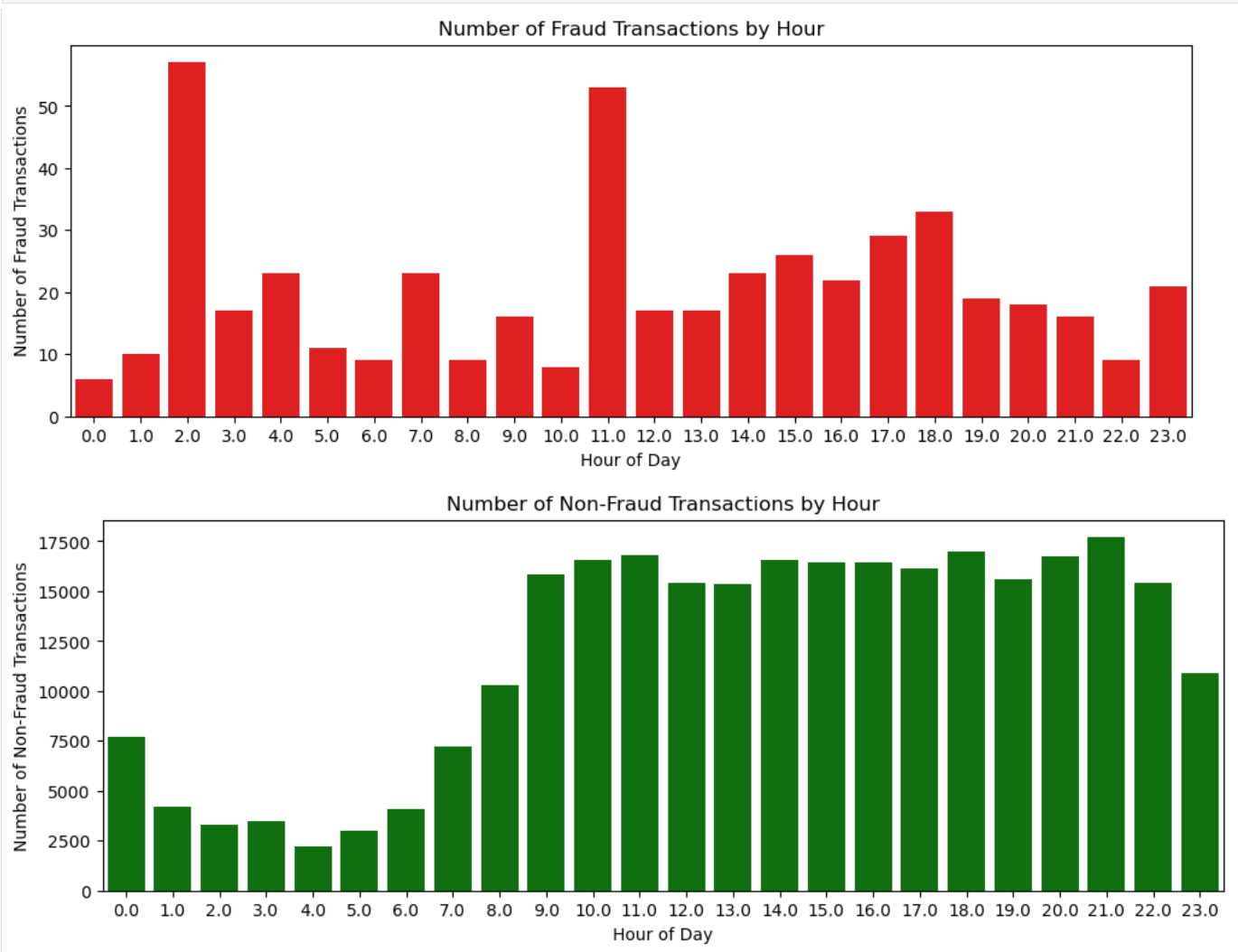
**💡 Keys:** Fraudulent transactions tend to be smaller in value and distributed evenly over time, making them easy to overlook if only looking at time or transaction amounts. This suggests that machine learning models are needed to detect fraud based on multiple features, rather than relying on simple visual inspection. These models can take into account various characteristics to more accurately identify unusual patterns and detect fraud that might be missed using basic methods.



The chart shows that **fraudulent transactions (Class = 1)** tend to have low values and a narrow distribution, while **legitimate transactions (Class = 0)** have a much broader range, including many transactions with very large amounts.

👉 Fraudulent transactions are typically smaller in value and exhibit less variation, whereas legitimate transactions vary widely in amounts. This characteristic can be exploited in fraud detection models.

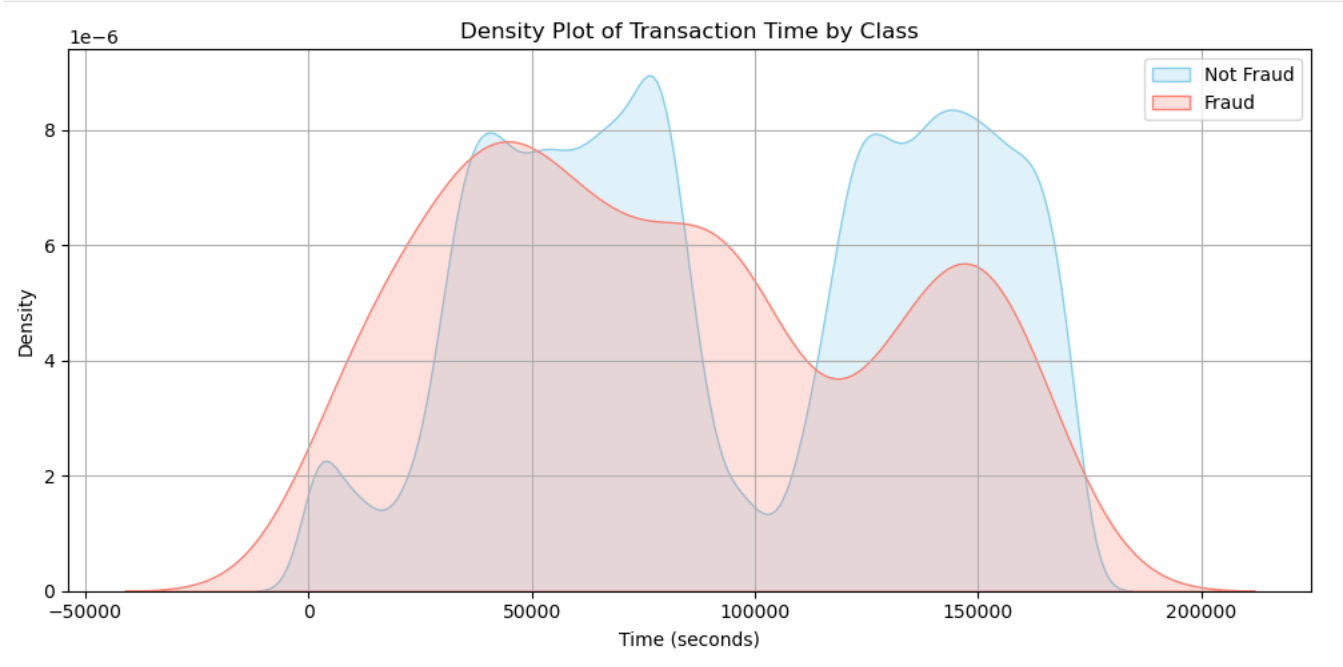
This aligns with the scatter plot, where fraudulent transactions are scattered randomly over time and rarely exceed a few thousand units. This observation reinforces the hypothesis that fraudsters intentionally split the transaction amounts into smaller values to avoid detection.



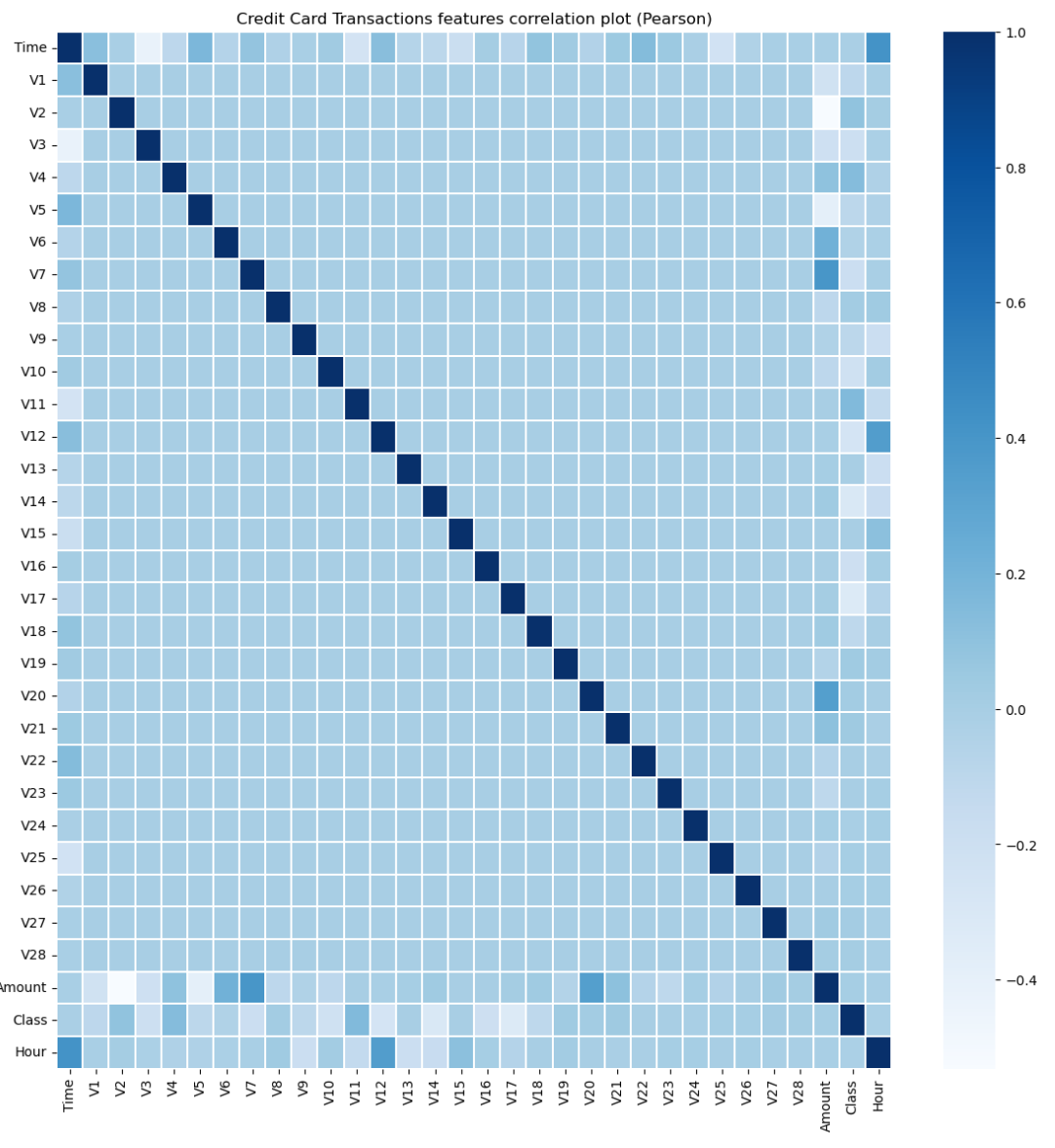
|  |  |  |
| --- | --- | --- |
| Characteristic | Fraudulent Transactions (Fraud) | Legitimate Transactions (Non-Fraud) |
| Volume | Very few (rarely > 1% of total transactions) | Majority (99%+) |
| Time Distribution | Evenly spread across 24 hours | Slight increase during the day, decreases at night |
| Average Value | Low (<$1,000) | Diverse (from $10 to $20,000+) |
| Variability | Low | High |

**Recommendations for Fraud Detection Model**

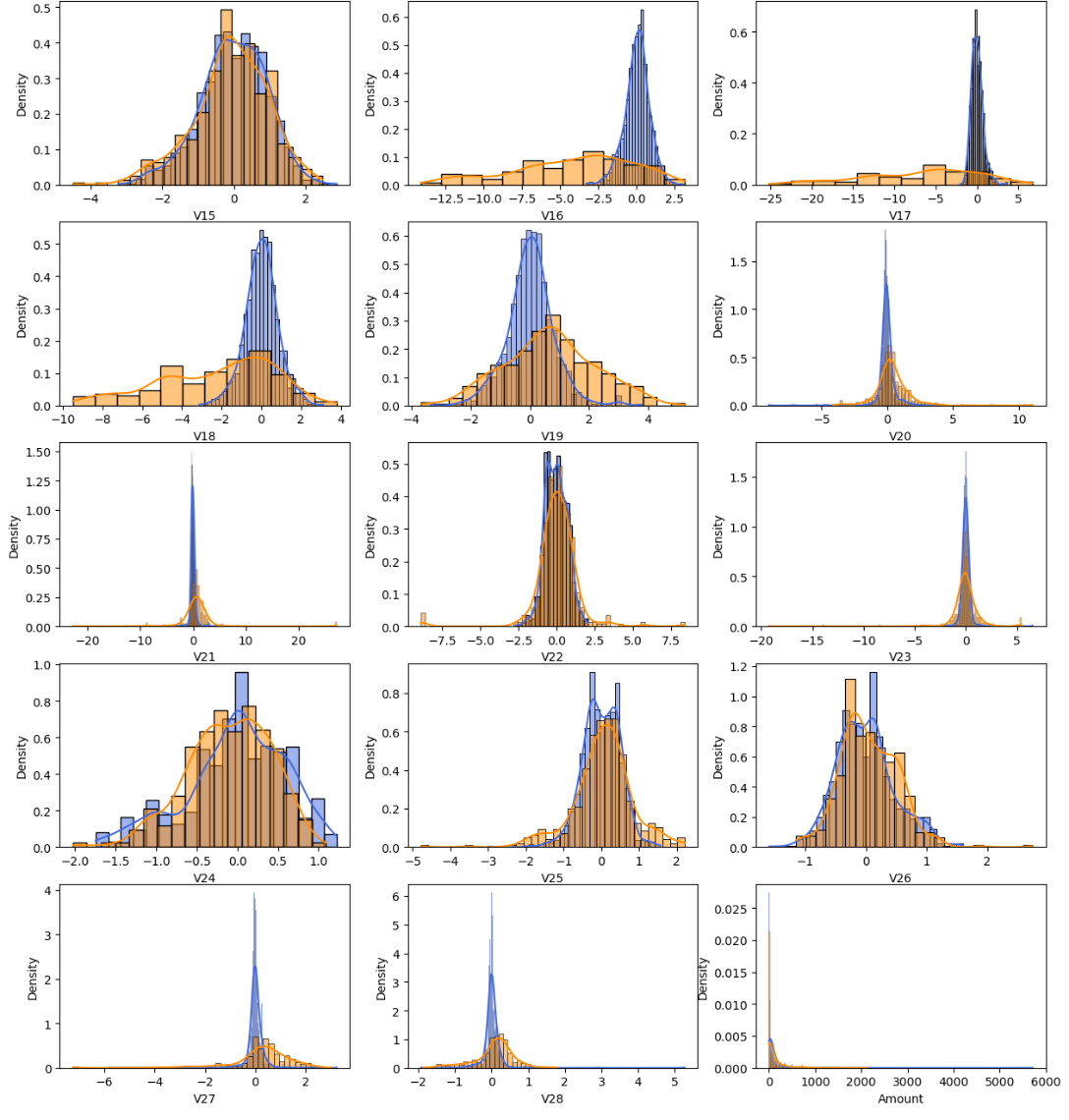
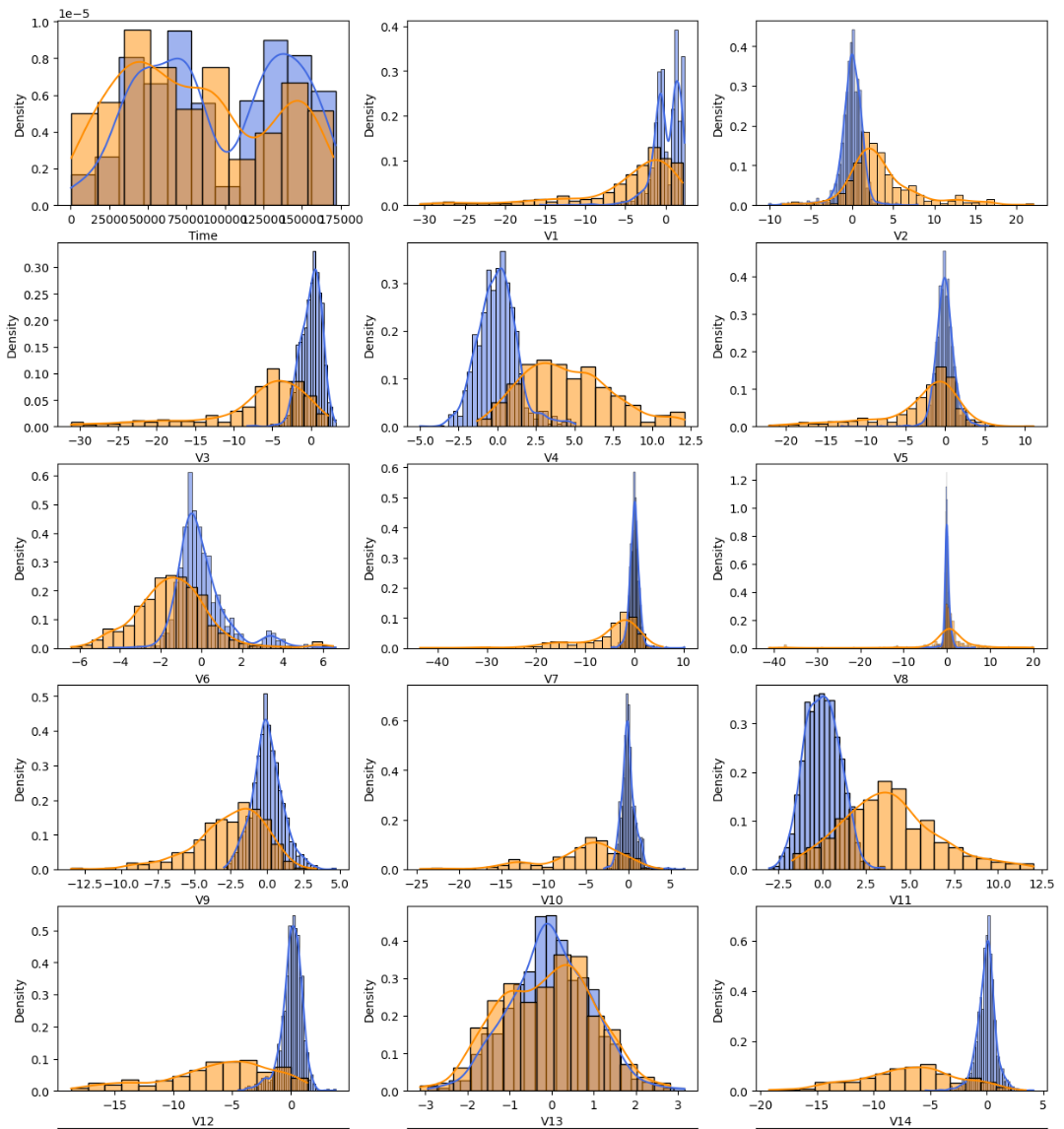
* **Leverage Time Features**:
  + Use Time-Series Analysis to detect anomalous transactions, even with small amounts (e.g., 5 transactions of $500 within 10 minutes from the same card).
* **Combine Multiple Features**:
  + Add features such as:
    - Transaction frequency per hour.
    - Deviation in value compared to the user's historical behavior.
    - Geographical location (e.g., if transactions are more than 100 km apart within one hour).
* **Handle Data Imbalance**:
  + Apply techniques like SMOTE or Class Weight adjustment in Random Forest models to prevent bias toward the majority class (Non-Fraud).
* **Dynamic Threshold Alerts**:
  + Set dynamic thresholds that vary based on time periods (e.g., set a threshold of 300 at night, and 500 during the day).



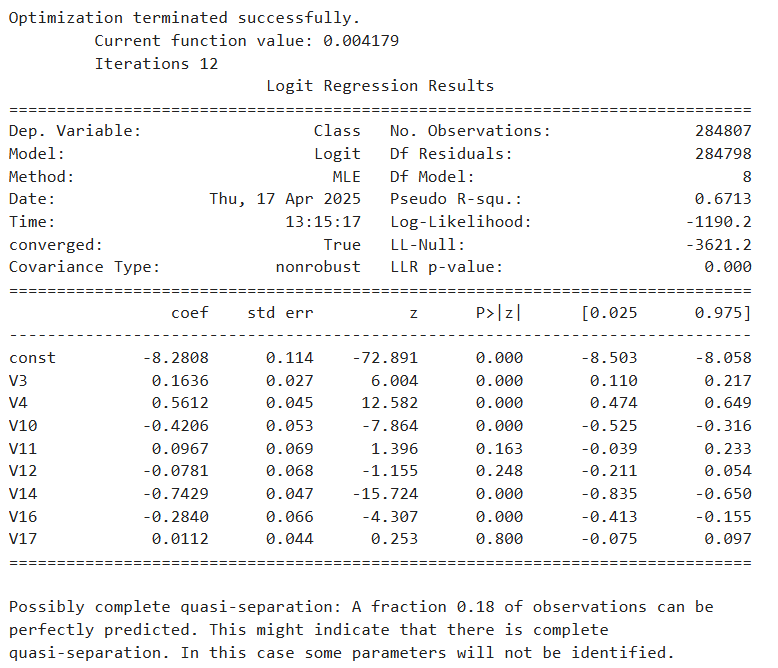
|  |  |  |
| --- | --- | --- |
| Feature | Legitimate Transactions | Fraudulent Transactions |
| Time Distribution | Clear peaks during certain hours (e.g., rush hours) | Evenly distributed, no prominent peaks |
| Density | High in specific time intervals | Low and stable across the timeline |
| Pattern/Trend | Follows human daily activity patterns | Random, possibly automated or bot-driven |



There is no notable correlation between features V1-V28. There are certain correlations between some of these features and Time (inverse correlation with V3) and Amount (direct correlation with V7 and V20, inverse correlation with V1 and V5).



For some of the features we can observe a good selectivity in terms of distribution for the two values of Class : V3, V4, V10, V11, V12, V14, V16, V17.

**The logistic regression results indicate that the model has converged successfully and demonstrates a relatively good fit, with a pseudo R-squared of 0.6713. Among the selected features, variables such as **V3**, **V4**, **V10**, **V14**, and **V16** are statistically significant with p-values less than 0.05, suggesting they contribute meaningfully to fraud detection. In particular, **V14** and **V10** exhibit strong effects, which aligns with prior domain insights.

However, some variables like **V11**, **V12**, and **V17** are not statistically significant, indicating limited explanatory power in this logistic setting. Moreover, the warning about possible quasi-separation — where a fraction of the data can be perfectly predicted — suggests potential overfitting or multicollinearity, which could limit model generalization.

This motivates the need to explore more robust or complex models, especially those suited for imbalanced classification problems. In the next steps, advanced machine learning method - **Random Forest** will be implemented and compared, aiming to improve performance in detecting fraudulent transactions.

# 4.3 Building the Random Forest Model

### 

**1. Data Preprocessing:**

* The Time column is removed as it is not necessary for the model.
* The target variable Class is separated from the feature set.
* Features are standardized using the StandardScaler method to ensure that all variables are on the same scale, improving model learning efficiency.

**2. Splitting the Dataset:**

* The dataset is split into training and testing sets, with 80% used for training and 20% for testing.
  + X\_train contains approximately **227,845** rows.
  + X\_test contains around **56,962** rows.
* Splitting is used to maintain the same ratio of fraud and non-fraud transactions in both sets.

**3. Addressing Class Imbalance:** Since fraudulent transactions are highly underrepresented, the **SMOTE** (Synthetic Minority Oversampling Technique) method is used to generate synthetic samples of the minority class, balancing the two classes before training.

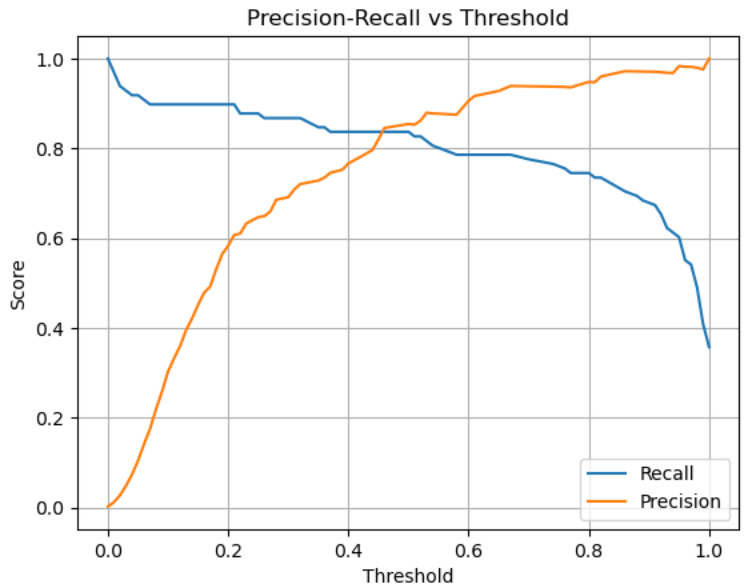
**4. Model Training:**

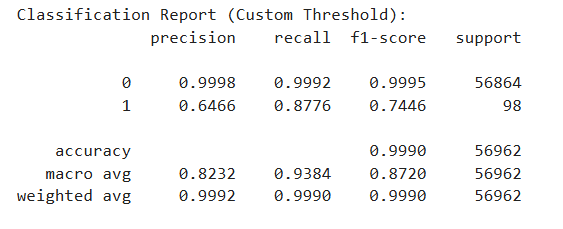
* A **Random Forest** model is employed due to its strength and interpretability in classification problems.
* The parameter class\_weight='balanced' is applied to prevent the model from being biased toward the majority class (legitimate transactions).

**5. Evaluation Using Custom Threshold:**

* A **Precision–Recall curve** is plotted across different thresholds to determine the most effective cutoff point.
* A custom threshold of **0.25** is selected instead of the default 0.5, aiming to balance high fraud detection (recall) while minimizing false positives (precision).

**4.4 Model Results**





**Class 0 – Legitimate Transactions (Class = 0)**

* **Precision = 99.98%** → Nearly all transactions predicted as legitimate were indeed legitimate.
* **Recall = 99.92%** → The model successfully identified almost all legitimate transactions in the test set.
* **F1-score = 99.95%** → The harmonic mean of precision and recall indicates near-perfect performance.
* This demonstrates that the model is extremely reliable in handling normal transactions, minimizing false positives (legitimate transactions incorrectly flagged as fraud).

**Class 1 – Fraudulent Transactions (Class = 1)**

* **Precision = 64.66%** → Of the transactions predicted to be fraudulent, about 65% were actually fraud. Around 35% were false alarms.
* **Recall = 87.76%** → The model was able to detect over 87% of actual fraudulent transactions. This is a strong point, as high recall is critical in fraud detection.
* **F1-score = 74.46%** → A relatively good overall performance, reflecting a balance between precision and recall.
* However, the lower precision indicates that the model sometimes incorrectly flags legitimate transactions as fraudulent. While this may cause inconvenience for users, it is often acceptable when prioritizing financial security.

**Overall Metrics**

* **Accuracy = 99.90%** → The model has a very high overall accuracy. However, due to class imbalance, this metric alone is not fully representative.
* **Macro average**: Evaluates both classes equally. Precision ~82%, Recall ~94%, F1 ~87% → showing the model performs well on both classes, especially fraud.
* **Weighted average**: Close to the accuracy score due to the overwhelming number of legitimate transactions.
  + **Conclusion:**
* The model effectively detects fraudulent transactions with a high recall (87.76%), which helps secure the financial system.
* Though the precision is moderate (64.66%), this is an acceptable trade-off to increase sensitivity in real-world scenarios.
* Future improvements may include optimizing the decision threshold, combining multiple models, or using manual review to reduce false positives.

### 5. Ethical Considerations

**5.1 Data Privacy and Anonymization**  
The dataset used in this project was anonymized using **PCA** transformation, with features labeled from V1 to V28. It contains no names, addresses, or personal identifiers, ensuring full compliance with privacy and data protection regulations.

**5.2 Transparency and Fairness in Modeling**  
The model was trained using transformed technical features, which helps avoid bias or discrimination. Evaluation is based on objective metrics such as precision, recall, and F1-score, promoting transparency and fairness in fraud detection.

### 6. Conclusion and Future Development

**6.1 Summary of Achievements**  
This project successfully implemented a pipeline to detect credit card fraud using real-world data. With over 280,000 transactions (only ~0.17% being fraud), the model underwent preprocessing, visualization, and data balancing via **SMOTE**. The **Random Forest** model, combined with a custom threshold, achieved impressive results:

* **Accuracy:** 99.90%
* **F1-score (Fraud):** 0.7446
* **Recall (Fraud):** 87.76% – strong ability to detect fraudulent activity
* **Precision (Fraud):** 64.66% – relatively accurate in identifying fraud  
  The analysis also revealed key insights – for example, fraud often appears disguised as small, randomly timed transactions.

**6.2 Limitations and Suggestions for Improvement**

**Limitations:**

* The model still misses some fraudulent transactions (false negatives), which poses real-world risks.
* Precision is not yet optimal → resulting in some false positives, affecting user experience.
* PCA-based anonymization limits the interpretability of individual features.

**Future Directions:**

* Implement advanced models such as **XGBoost**, **LightGBM**, or ensemble approaches.
* Optimize the classification threshold based on business goals (e.g., cost of fraud vs. cost of manual review).
* Enrich the dataset with external information (e.g., transaction location, device data, user behavior patterns).

# Appendix

# Code Python: <https://github.com/nguyenhuyen2004/creditcard/blob/master/creditcard.ipynb>