Credit Card Fraud Detection

Submitted by

Samiksha Band
Bhawana Arunrao Jirapure
Parag Mahendrakumar Buch
Kavita Lamani

Date: 25th May 2024

Abstract:

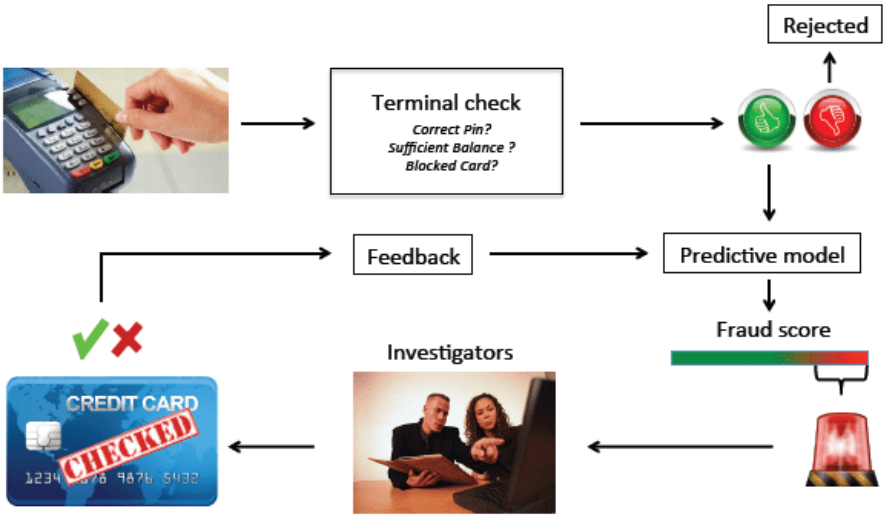
Credit card fraud refers to the physical loss of credit card or loss of sensitive credit card information. This product report outlines the development of a credit card fraud detection and prevention system employing machine learning techniques. The report addresses the escalating issue of credit card fraud globally and presents a comprehensive approach to developing an advanced system capable of real-time fraud detection and prevention while meeting the needs of financial institutions and consumers. In present scenario when the term fraud comes into a discussion, credit card fraud clicks to mind so far. With the great increase in credit card transactions, credit card fraud has increasing excessively in recent years. Fraud detection includes monitoring of the spending behavior of users/ customers in order to determination, detection, or avoidance of undesirable behavior. As credit card becomes the most prevailing mode of payment for both online as well as regular purchase, fraud relate with it are also accelerating. This report shows Random Forest algoriths that can be used for classifying transactions as fraud or genuine one. The work is implemented in Python.
Keywords- Credit Card Fraud Detection, Credit card frauds, Fraud detection system, fraud detection, real-time credit card fraud detection, skewed distribution.

Introduction:

Credit card is a small thin plastic or fiber card that contains information about the person such as picture or signature and person named on it to charge purchases and service to his linked account charges for which will be debited regularly. Now a day's card information is read by ATM's, swiping machines, store readers, bank and online transaction. Each card has a unique card number which is very important, its security is mainly relies on physical security of the card and also privacy of the credit card number. There is rapid increase in the credit card transaction which has led to substantial growth in fraudulent cases. Detection of fraud using efficient and secure methods is very important.

When we make any transaction while purchasing any product online, a good amount of people prefers credit cards. The credit limit in credit cards sometimes helps us in making purchases even if we don't have the amount at that time. but, on the other hand, these features are misused by cyber attackers. To tackle this problem, we need a system that can abort the transaction if it finds fishy. Here, comes the need for a system that can track the pattern of all the transactions and if any pattern is abnormal then the transaction should be aborted. Today, we have many machine learning algorithms that can help us classify abnormal transactions. The only requirement is the past data and the suitable algorithm that can fit our data in a better form.

There are growing numbers of new companies all around the world. All of those companies are trying to provide best service quality for their customers. In order to succeed in that, companies are processing a lot of data on a daily basis. These data come from vast number of resources and are in different formats. Moreover, this data contains some of the key parts of the company's future business. Because of that, companies need to store that data, to process it and what is really important, to keep it safe. Without securing data, a lot of it can be used by other companies or even worse, it can be stolen. In most cases, financial information is stolen, which can harm whole company or individual.



Problem Statement:

Credit card fraud has become a pervasive issue, causing significant financial losses to businesses and consumers alike. Traditional fraud detection methods are no longer sufficient to combat increasingly sophisticated fraudulent activities. There is an urgent need for a robust system that can accurately detect and prevent fraudulent transactions in real-time to safeguard financial transactions and mitigate losses.

Objective:

The objective of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction.

Data Sources:

Dataset

In this Credit Card Fraud Detection dataset is used, which is downloaded from Kaggle [17]. This dataset contains transactions, occurred in two days, made in September 2013 by European cardholders. The dataset contains 31 numerical features. Since some of the input variables contains financial information, the PCA transformation of these input variables were performed in order to keep these data anonymous. Three of the given features weren't transformed. Feature "Time" shows the time between first transaction and every other transaction in the dataset. Feature "Amount" is the amount of the transactions made by credit card. Feature "Class" represents the label, and takes only 2 values: value 1 in case of fraud transaction and 0 otherwise.

Dataset contains 2,18,660 transactions where 469 transactions were frauds and the rest were genuine. Considering the numbers, we can see that this dataset is highly imbalanced, where only 0.20 % of transactions are labeled as frauds.

Data Gathering:

	id	Time	V1	V2	V3	V4	V5	V6	V7	V8	...	V21	V22	V23	V24	V25	
0	0	0.0	2.074329	-0.129425	-1.137418	0.412846	-0.192638	-1.210144	0.110697	-0.263477	...	-0.334701	-0.887840	0.336701	-0.110835	-0.291459	0.20
1	1	0.0	1.998827	-1.250891	-0.520969	-0.894539	-1.122528	-0.270866	-1.029289	0.050198	...	0.054848	-0.038367	0.133518	-0.461928	-0.465491	-0.46
2	2	0.0	0.091535	1.004517	-0.223445	-0.435249	0.667548	-0.988351	0.948146	-0.084789	...	-0.326725	-0.803736	0.154495	0.951233	-0.506919	0.08
3	3	0.0	1.979649	-0.184949	-1.064206	0.120125	-0.215238	-0.648829	-0.087826	-0.035367	...	-0.095514	-0.079792	0.167701	-0.042939	0.000799	-0.09
4	4	0.0	1.025898	-0.171827	1.203717	1.243900	-0.636572	1.099074	-0.938651	0.569239	...	0.099157	0.608908	0.027901	-0.262813	0.257834	-0.25

5 rows × 32 columns

```
1 df_train.shape
```

```
(219129, 32)
```

Exploring Data:

A First Glimpse at the Data:

- After data collection, exploratory data analysis cleans and – if necessary – preprocesses the data.
- This exploration stage also offers guidance on the most suitable algorithm for extracting meaningful market segments.
- At a more technical level, data exploration helps to (1) identify the measurement levels of the variables; (2) investigate the univariate distributions of each of the variables; and (3) assess dependency structures between variables. In addition, data may need to be pre-processed and prepared so it can be used as input for different segmentation algorithms.
- Results from the data exploration stage provide insights into the suitability of different segmentation methods for extracting market segments.

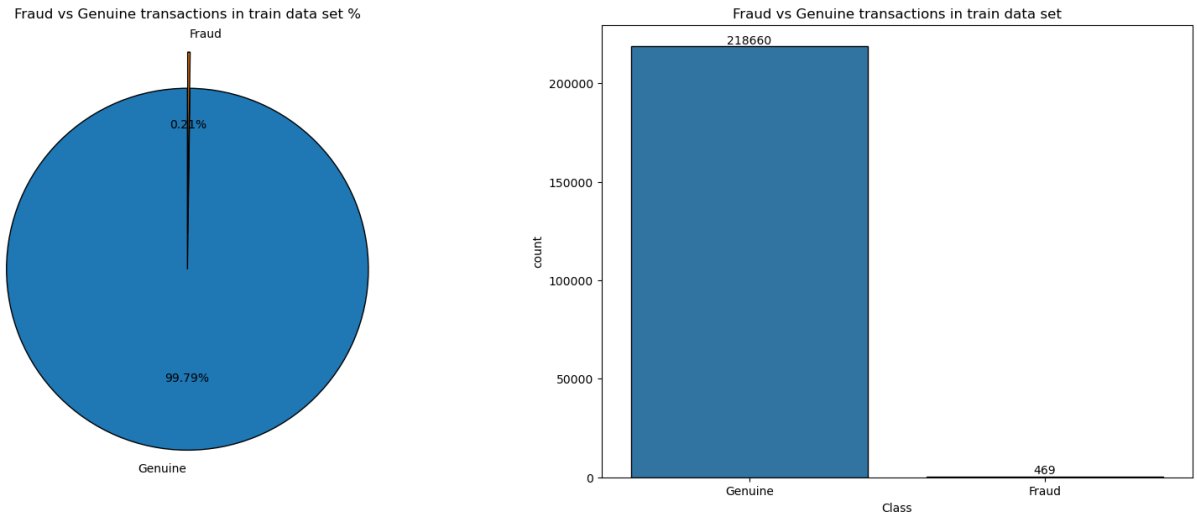


Fig. Fraud vs Genuine transactions in train dataset

The pie chart shows the percentage of genuine transactions versus fraudulent transactions. In this case, the values are highly imbalanced (which is common in fraud detection datasets), it visually highlights this imbalance. 99.8% of transactions are genuine and only 0.2% are fraudulent.

This bar plot (count plot) shows the count of genuine and fraudulent transactions in the training dataset. There are 2,18,660 genuine transactions and 469 fraudulent transactions.

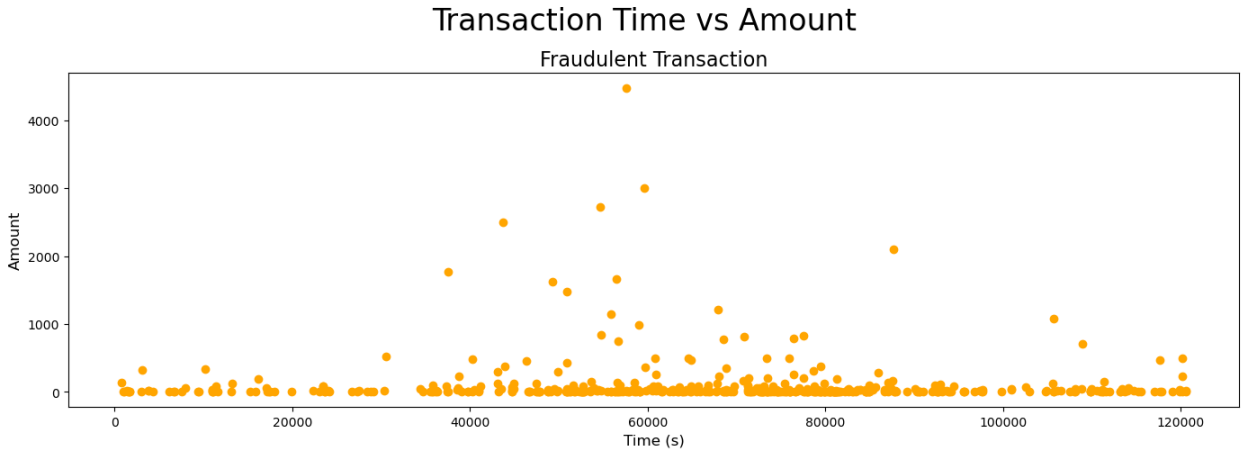


Fig. Transaction Time vs Amount for fraudulent transaction

This plot reveals the distribution and pattern of fraudulent transactions over time and their corresponding amounts.

It helps in identifying any clustering or trends specific to fraudulent transactions, such as certain times of day or transaction amounts where fraud is more prevalent.

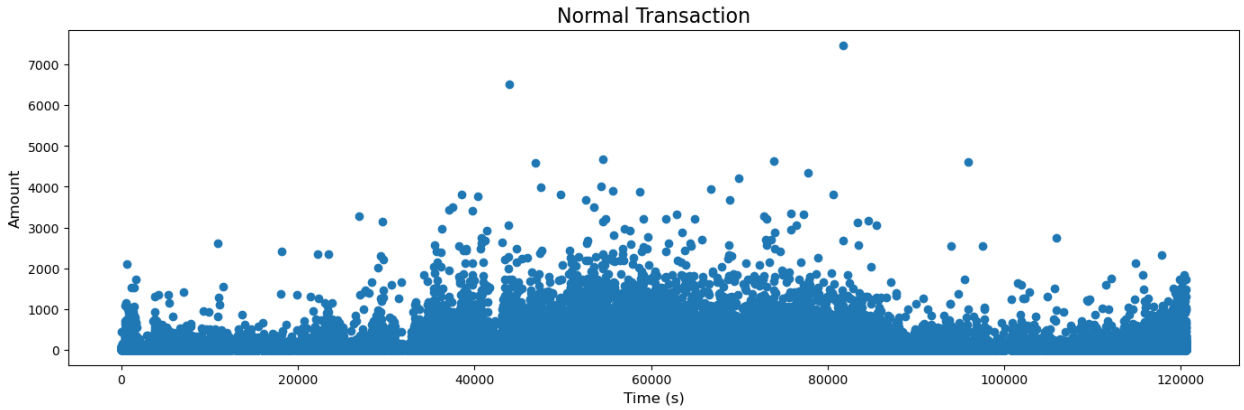


Fig. Transaction Time vs Amount for normal transaction

This plot shows the distribution and pattern of normal transactions over time and their amounts.

It provides a baseline to compare against fraudulent transactions, helping to identify any anomalous patterns or deviations that are indicative of fraud.

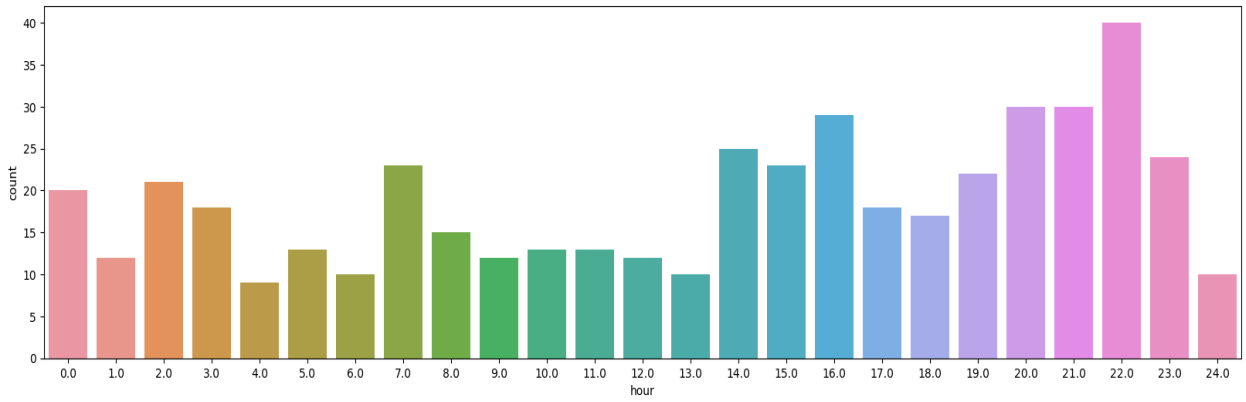


Fig. Hourly count

the plot shows a significant number of fraudulent transactions around 4 PM and 10 PM, it suggests that fraudsters might be targeting these hours, possibly due to lower monitoring and security measures in place during off-peak hours.

Conversely, hours with low counts of fraudulent transactions might indicate times when fraud attempts are less frequent, possibly due to higher vigilance or lower transaction volumes.

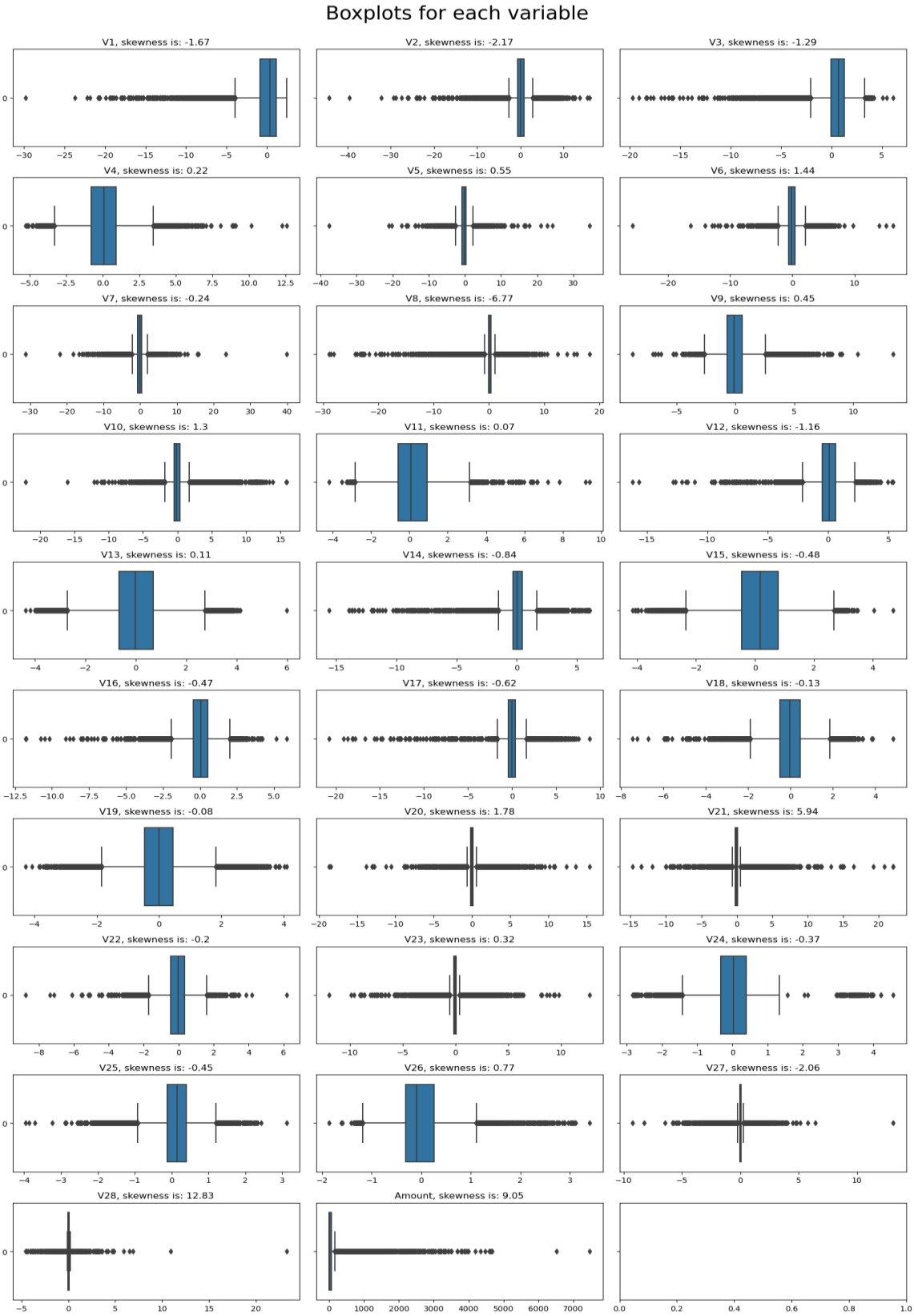


Fig. Boxplot for each variable

Correlation Heatmap:

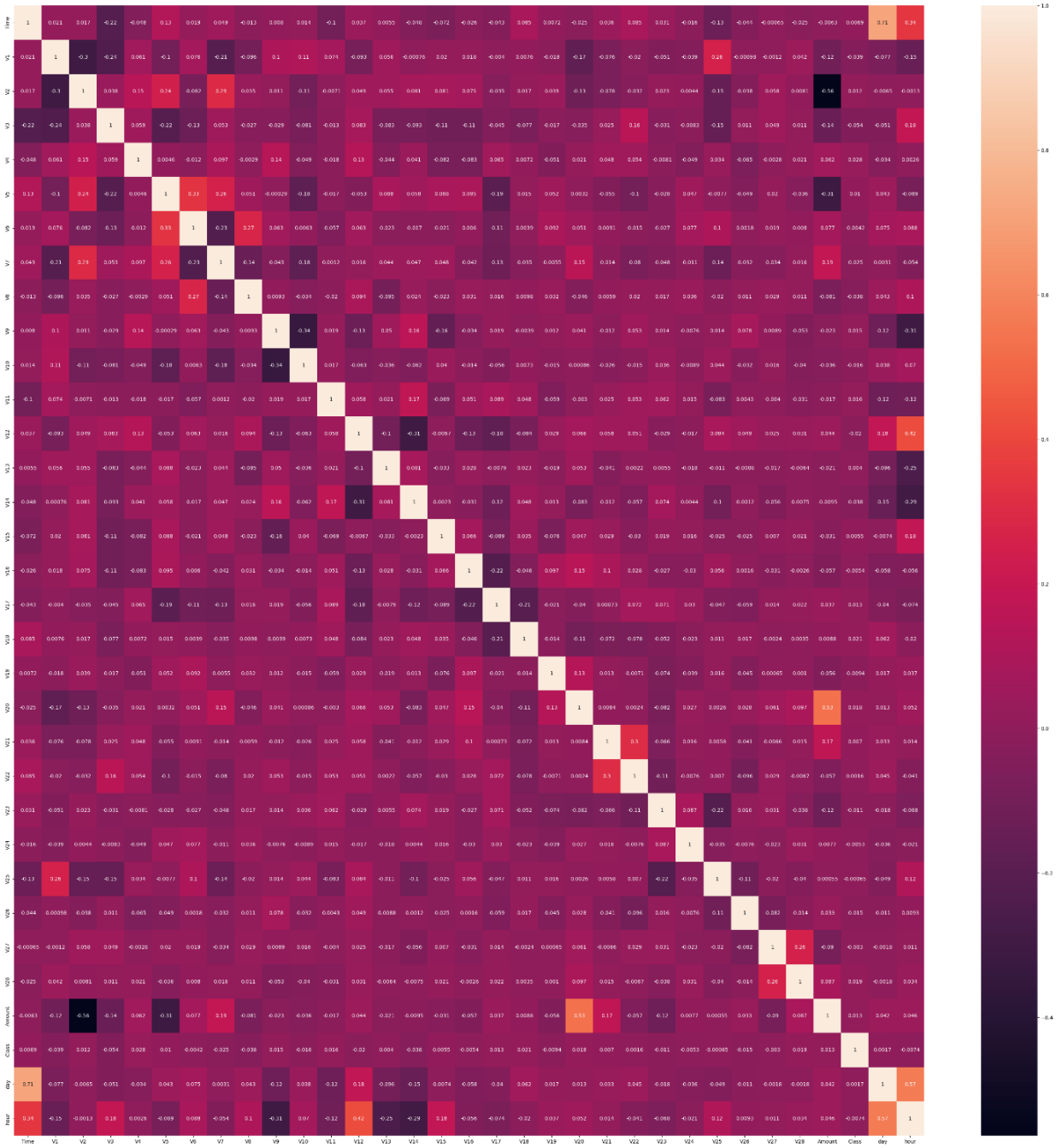


Fig. Correlation Heatmap

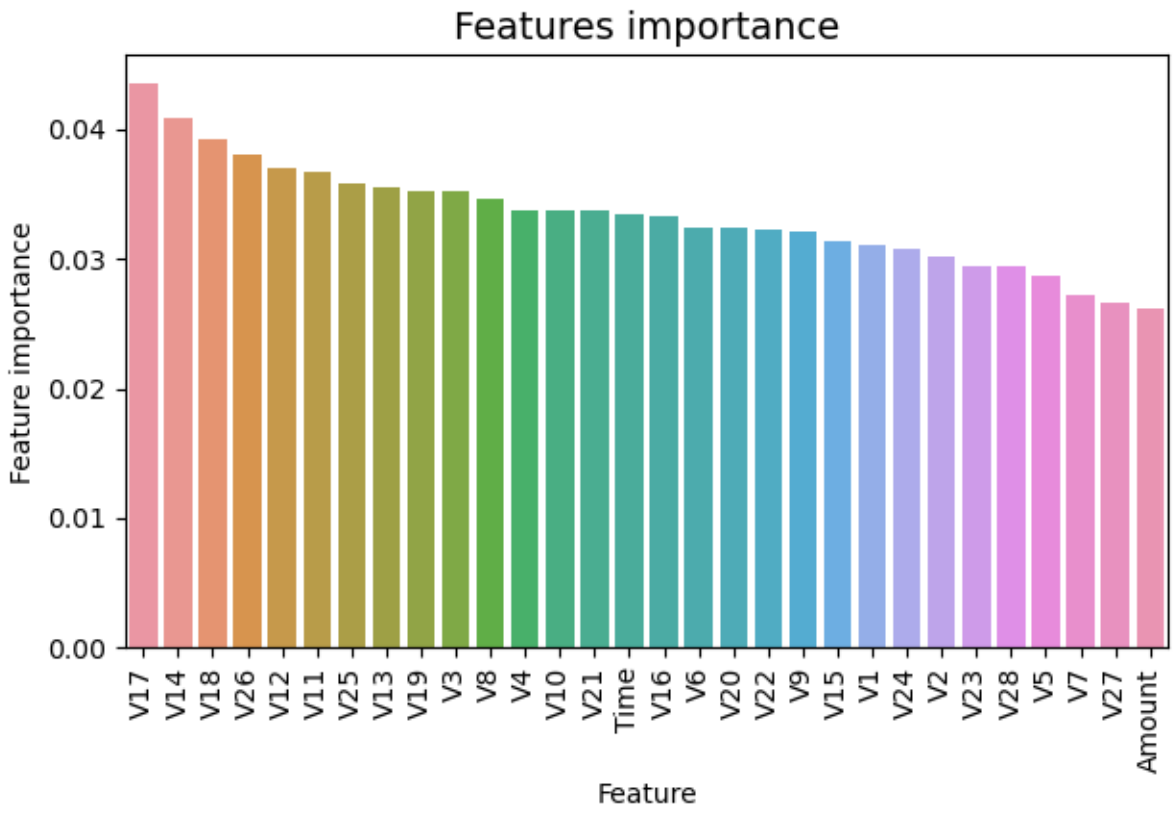


Fig. Important features of dataset

Confusion Matrix:

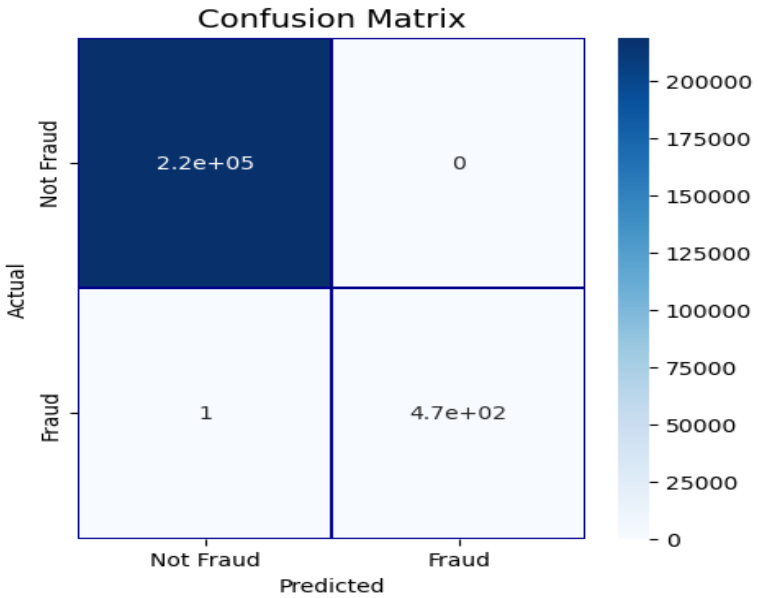


Fig. Confusion Matrix

Area under curve:

```
1 roc_auc_score(df_train[target].values, preds)
```

```
0.9989339019189765
```

An ROC AUC score of 0.99 implies that the model is able to perfectly distinguish between the positive and negative classes in almost all cases. The model's predictions are very accurate, with minimal misclassifications.

The high ROC AUC score suggests that the classifier has both high sensitivity (true positive rate) and high specificity (true negative rate). It means the model effectively identifies true positives while minimizing false positives.

A score of 0.99 indicates that the model's performance is consistent across different thresholds. In other words, even when you adjust the decision threshold for classification, the model maintains its high level of performance.

Business Modelling:

1. Data Collection:

- Transaction data
- User behavior data
- Device information
- Geolocation data
- Merchant data
- Historical fraud data

2. Data Processing:

- Data cleansing
- Data normalization
- Feature extraction
- Data integration

3. Machine Learning Models:

- Anomaly detection
- Pattern recognition

- Behavioral analysis
- Predictive modeling

4. Real-time Monitoring:

- Monitoring transactions
- Applying fraud rules
- Threshold-based alerts
- Transaction stream analysis

5. User Profiling:

- Creating user profiles
- Analyzing spending patterns
- Identifying usual behavior
- Customized risk assessment

6. Biometric Authentication:

- Fingerprint scanning
- Facial recognition
- Voice recognition
- Two-factor authentication

7. Transaction Verification:

- SMS alerts
- Push notifications
- In-app prompts
- User confirmation/denial

8. Fraud Response:

- Blocking suspicious transactions
- Temporarily freezing accounts
- Notifying users and issuers
- Fraud investigation workflows

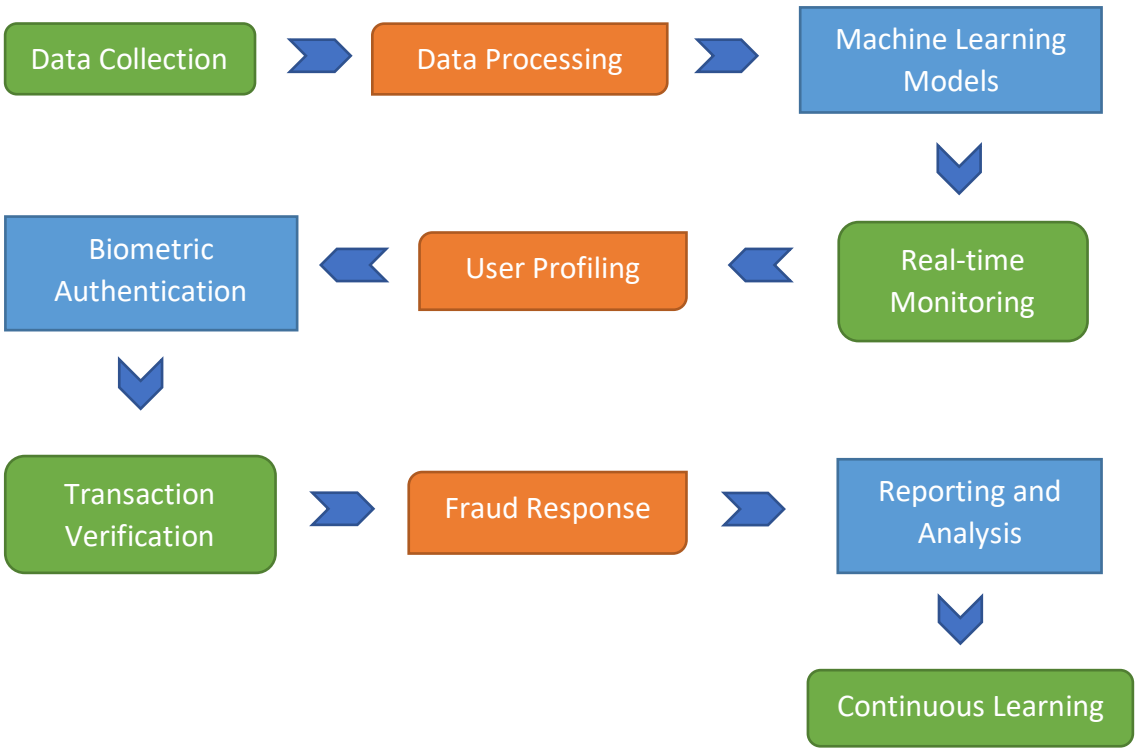
9. Reporting and Analysis:

- Generating fraud reports
- Analyzing fraud incidents

- Identifying fraud patterns
- Improving fraud prevention strategies

10. Continuous Learning:

- Updating machine learning models
- Incorporating new data
- Adapting to emerging fraud tactics
- Enhancing detection accuracy



Financial Modelling:

1. Assumptions

- Number of Transactions Monitored (per month): 1,000,000
- Fraud Detection Rate: 98%
- False Positive Rate: 1%
- Average Transaction Value: ₹10,000
- Fraudulent Transaction Value (per transaction): ₹10,000
- Cost Savings per Detected Fraud: ₹9,000
- Marketing Costs: ₹70,000
- Salaries: ₹300,000
- Technology Infrastructure Costs: ₹150,000
- Maintenance Costs: ₹50,000
- Number of Clients Using the System: 10
- Subscription Fee per Client (monthly): ₹50,000
- Setup Fee per Client: ₹100,000

2. Revenue Streams

1. Subscription Fees: Monthly fees charged to clients for using the fraud detection system.
    - Subscription Revenue: Number of Clients x Subscription Fee per Client x 12 months
    - Subscription Revenue=10×₹50,000×12=₹6,000,000
    - Subscription Revenue=10×₹50,000×12=₹6,000,000
2. Setup Fees: One-time setup fees charged to new clients.
    - Setup Revenue: Number of Clients x Setup Fee per Client
    - Setup Revenue=10×₹100,000=₹1,000,000
3. Cost Savings from Fraud Detection: Savings accrued by preventing fraudulent transactions.
    - Total Transactions Detected as Fraudulent: Number of Transactions Monitored x Fraud Detection Rate x Fraudulent Transaction Rate
    - Total Fraudulent Transactions Detected=1,000,000×0.98×0.001=980
    - Total Savings: Total Fraudulent Transactions Detected x Cost Savings per Detected Fraud

- Total Savings=980×₹9,000=₹8,820,000

3. Total Revenue

- Total Revenue: Subscription Revenue + Setup Revenue + Total Savings
- Total Revenue=₹6,000,000+₹1,000,000+₹8,820,000=₹15,820,000

4. Operating Expenses

1. Marketing Costs: ₹70,000
2. Salaries: ₹300,000
3. Technology Infrastructure Costs: ₹150,000
4. Maintenance Costs: ₹50,000

- Total Operating Expenses: Marketing Costs + Salaries + Technology Infrastructure Costs + Maintenance Costs
- Total Operating Expenses=₹70,000+₹300,000+₹150,000+₹50,000=₹570,000

5. Net Profit

- Net Profit: Total Revenue - Total Operating Expenses
- Net Profit=₹15,820,000−₹570,000=₹15,250,000

Summary

- Total Revenue: ₹15,820,000
- Total Operating Expenses: ₹570,000
- Net Profit: ₹15,250,000

Website Prototype:

https://fraud-guard-m5i5mac.gamma.site/

GitHub Link:

https://github.com/Samikshaband/credit-card-fraud-detection