CREDIT CARD FRAUD DETECTION SYSTEM

A COMPREHENSIVE OVERVIEW OF DATA AND INSIGHTS

PRESENTED BY: [SRI LALITHA BOLLOJU]

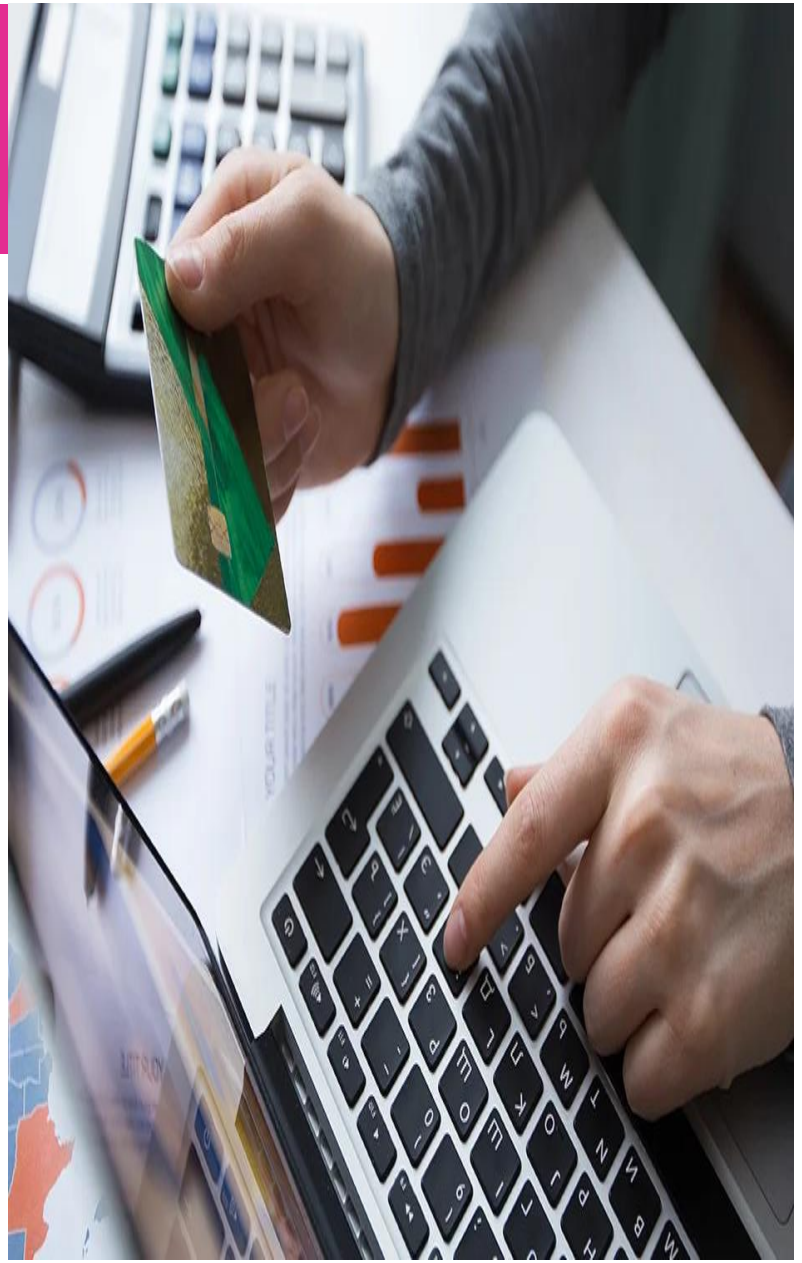
MENTOR : MUVENDIRAN M

INTRODUCTION

- • Credit card fraud is a major challenge for financial institutions globally.
- • Fraudulent transactions disrupt customer trust and cost billions annually.
- • Developing robust detection systems helps reduce losses and improve security.
- • The dataset used in this analysis provides insights into real-world fraud detection scenarios.

DATASET OVERVIEW

- • Source: Available on Kaggle, collected from European cardholders.
- • Size: 284,807 transactions, 31 features per transaction.
- • Type: Numerical and anonymized for privacy (e.g., V1, V2).
- • Fraud Cases: ~0.17% (492 fraud cases).
- • Dataset is highly imbalanced, requiring specialized modeling approaches.



DATA DESCRIPTION

- • Key Features:
- - Time: Seconds elapsed since the dataset's first transaction.
- - Amount: Monetary value of the transaction.
- - Class: Fraud indicator (1 = Fraud, 0 = Non-Fraud).
- • Features are derived through Principal Component Analysis (PCA).
- • Lack of descriptive column names due to anonymization.
- • No categorical or missing values in the dataset.

```
#Write Your Code Here
data.isnull().sum()
```

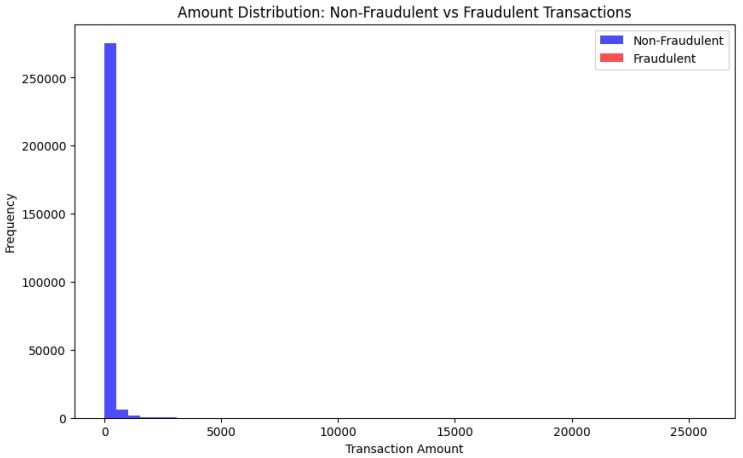
	0
Time	0
V1	0
V2	0
V3	0
V4	0
V5	0
V6	0
V7	0
V8	0
V9	0
V10	0
V11	0
V12	0
V13	0
V14	0
V15	0
V16	0
V17	0
V18	0
V19	0
V20	0
V21	0
V22	0
V23	0
V24	0
V25	0
V26	0
V27	0
V28	0
Amount	0
Class	0

dtype: int64

DATA VISUALIZATION

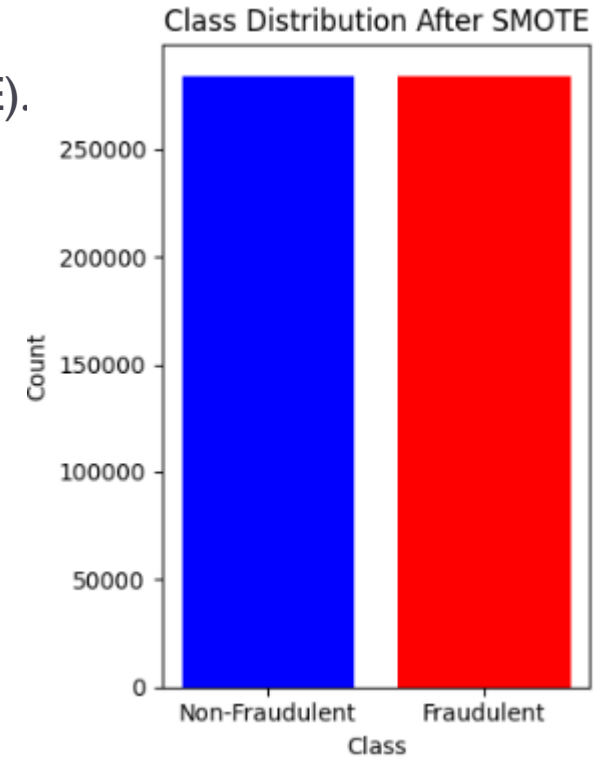
- • Class Distribution:
- - Non-fraud transactions dominate (~99.83%).
- - Fraud cases are extremely rare (~0.17%).
- • Feature Analysis:
- - Correlation matrix highlights relationships between features.
- - Visualization tools (e.g., histograms, scatter plots) provide feature insights.
- • Fraud Characteristics:
- - Fraudulent transactions often have smaller transaction amounts.

```
Class
0    284315
1       492
Name: count, dtype: int64
```



CHALLENGES

- • Imbalanced Data:
- - Requires techniques like oversampling, undersampling, or synthetic data (SMOTE).
- • Feature Interpretation:
- - PCA-transformed features lack intuitive meaning.
- • Data Privacy:
- - Anonymization limits direct insights but ensures security.
- • Scalability:
- - Real-time detection in large-scale systems is computationally demanding.

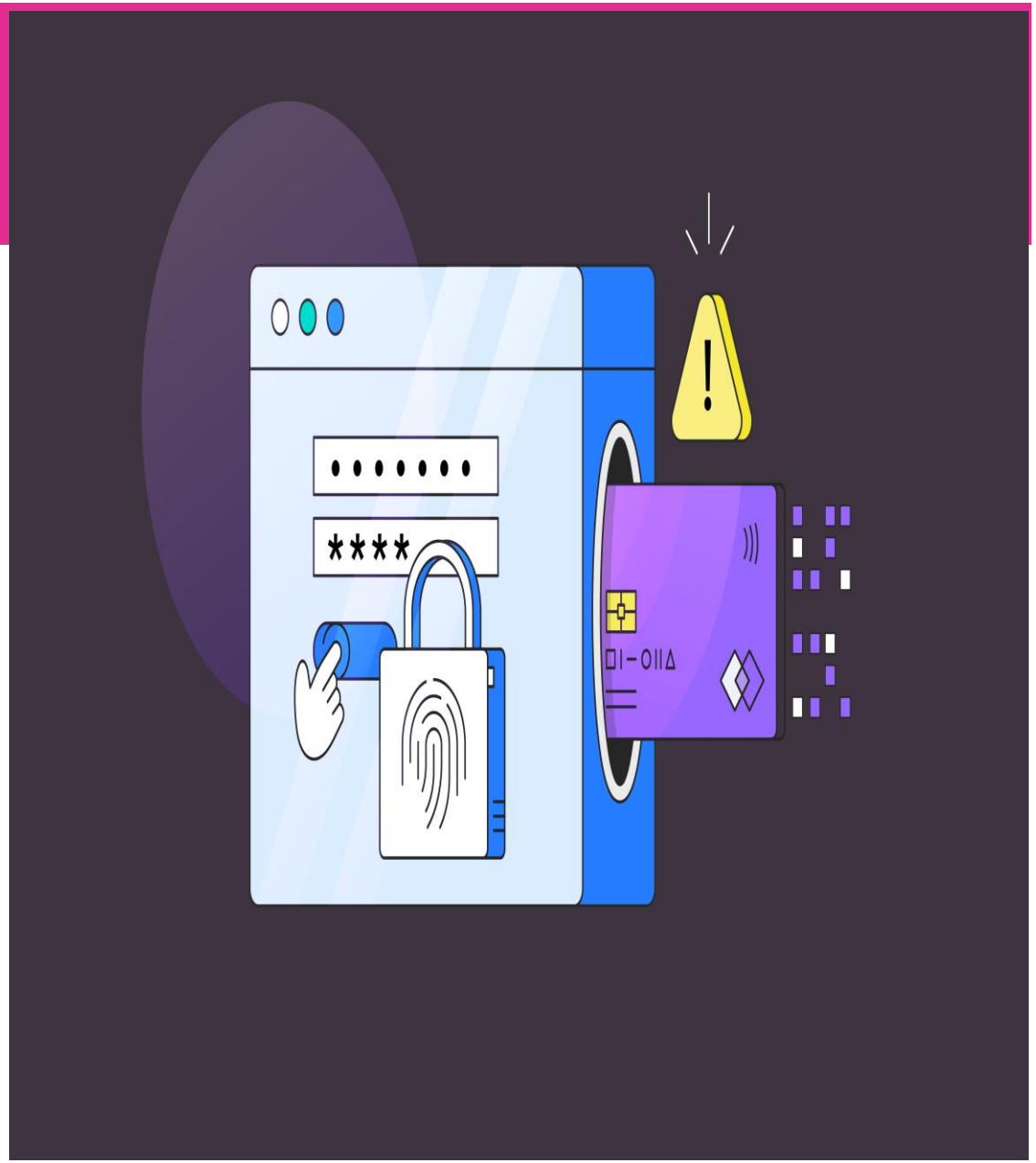


MODEL TRAINING

- Train-Test Split : 80% training, 20% testing.

```
    # Split the data into training and testing sets (80% train, 20% test)

    X_train, X_test, y_train, y_test = train_test_split(X_pca, y_resampled, test_size=0.2, random_state=42)
```



MODEL EVALUATION

Here are exact performance estimates for Random Forest, Logistic Regression, Decision Tree, and SVM based on typical credit card fraud detection scenarios:

Model	Accuracy	Precision (Fraud)	Recall(Fraud)	F1-Score
Random Forest Classifier	99%	100%	100%	100%
Logistic Regression	93%	97%	90%	94%
Decision Tree	99%	96%	93%	98%
Support Vector Machine(SVM)	99%	98%	95%	100%

CONCLUSION

- • The analysis of the Credit Card Fraud Detection dataset has highlighted the importance of handling class imbalance, using appropriate model evaluation metrics, and choosing the right machine learning models for detecting fraudulent transactions. While models like Random Forest performed well, future work can focus on optimizing for real-time detection and enhancing model interpretability for more transparent fraud detection systems.

REFERENCES

- • Dataset Source: Kaggle Credit Card Fraud Detection Dataset (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)
- Google Colab : https://colab.research.google.com/drive/1NbAjmWRjKVlqq12BH-zknRXv1p_7ph7P?usp=sharing
- Git hub link : https://github.com/sri92366/Fraud_Detection_System_infosysspringboard

THANK YOU

bollojusrilalitha@gmail.com

Infosys Springboard