Credit Card Fraud Detection Analysis

By:

Sweekrit Acharya

Charmi Raghavani

Reebika Bhatta

Shirish Thapaliya

Ruchita Soni

Overview

- Introduction
- Target Audience
- Problem and Needs
- Data Overview
- Data Preprocessing
- Methodology
- Analysis and Results
- Conclusion
- Reference

Introduction

Credit card fraud detection is the process of identifying and preventing unauthorized or fraudulent transactions made using credit cards, typically employing algorithms and machine learning models to analyze patterns and anomalies in transaction data [1].

Primary Goal: Creating an intelligent model that can accurately differentiate between genuine and fraudulent transactions in real-time.

Importance: Safeguarding consumers and financial institutions from financial losses, maintaining trust in the banking system, and fostering a secure environment for digital commerce.

Target Audience

Financial Institutions

Provide Financial Institutions with advanced tools and techniques to detect and prevent fraudulent activities, thereby minimizing financial losses and preserving the trust of their customers.

Consumers

Consumers are directly benefited by our project as it enhances the security of their credit card transactions, protecting them from financial losses and potential identity theft.

Business

Businesses that rely on electronic transactions benefit indirectly from our project as it contributes to a safer and more secure environment for conducting online commerce, thereby reducing the risk of financial losses due to fraudulent activities.

Possible Challenges:

- Sophisticated Fraud Techniques: Fraudsters continually evolve their tactics to bypass detection systems, requiring detection methods to stay ahead of emerging fraud patterns.
- Imbalanced Data: Fraudulent transactions are typically rare compared to legitimate ones, leading to imbalanced datasets that can skew model performance and accuracy[1].
- Real-time Processing: The need for timely detection and prevention of fraud requires systems capable of processing large volumes of transactions in real-time. Real-world credit card transaction datasets cannot be publicly available because they contain sensitive information about customers[3].
- Model Interpretability: Understanding how fraud detection models make decisions is crucial for trust and regulatory compliance, highlighting the need for interpretable machine learning models.

Why we need Credit Card Fraud Detection?

- Early Detection: Predicts fraudulent activities swiftly, preventing significant financial losses.
- Adaptability: Evolve with fraud patterns, detecting new tactics effectively.
- Accuracy: Identifies subtle anomalies, enhancing detection precision over rule-based systems.
- Enhanced Customer Experience: Minimizes impacts on legitimate transactions, improving customer satisfaction.

Dataset Overview

Data Source: Kaggle

1. The dataset comprises 284,808 records of historical credit card transactions.

2. The dataset consists of several features providing insights into each transaction, including:
   - **Time:** Timestamp indicating the date and time of the transaction
   - **Amount:** The monetary value of the transaction.
   - **V1-V28:** Principal components resulting from PCA transformation for anonymization purposes.
   - **Class**: Binary label indicating the transaction's legitimacy, with 0 representing legitimate transactions and 1 representing fraudulent transactions.

Data Preparation

Data Cleaning

- Addressing missing values and outliers.
- The original dataset had 284,807 rows, after removing duplicates 283,726 unique rows in the dataset left.

Data Transformation

- This phase involves techniques like normalization, standardization, encoding categorical variables, and feature engineering to improve model predictions.

Transaction Amount Binning

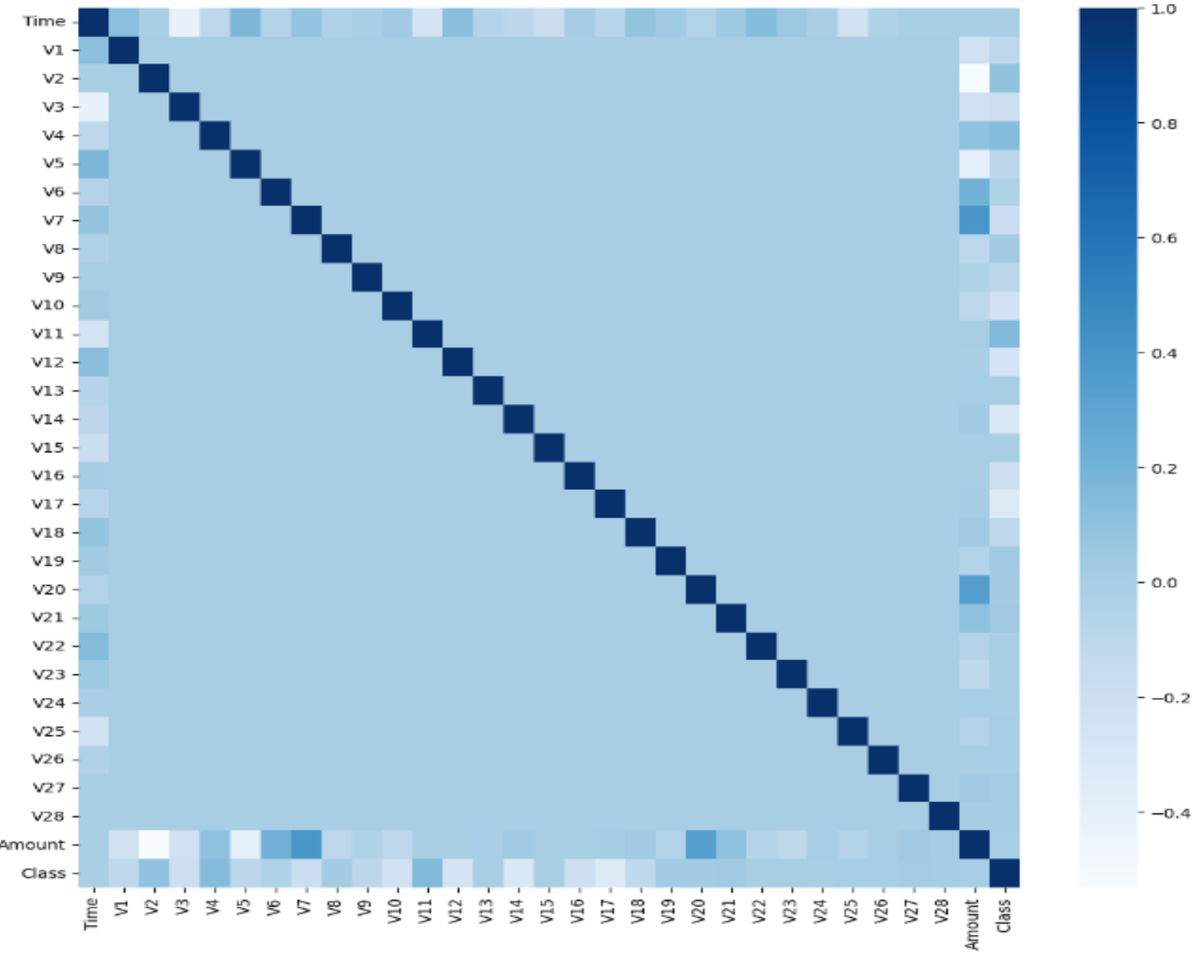
- Binning transaction amounts into "Amount_Category" feature can enhance model generalization.
- Categorization helps mitigate variability and noise in raw transaction data.

.

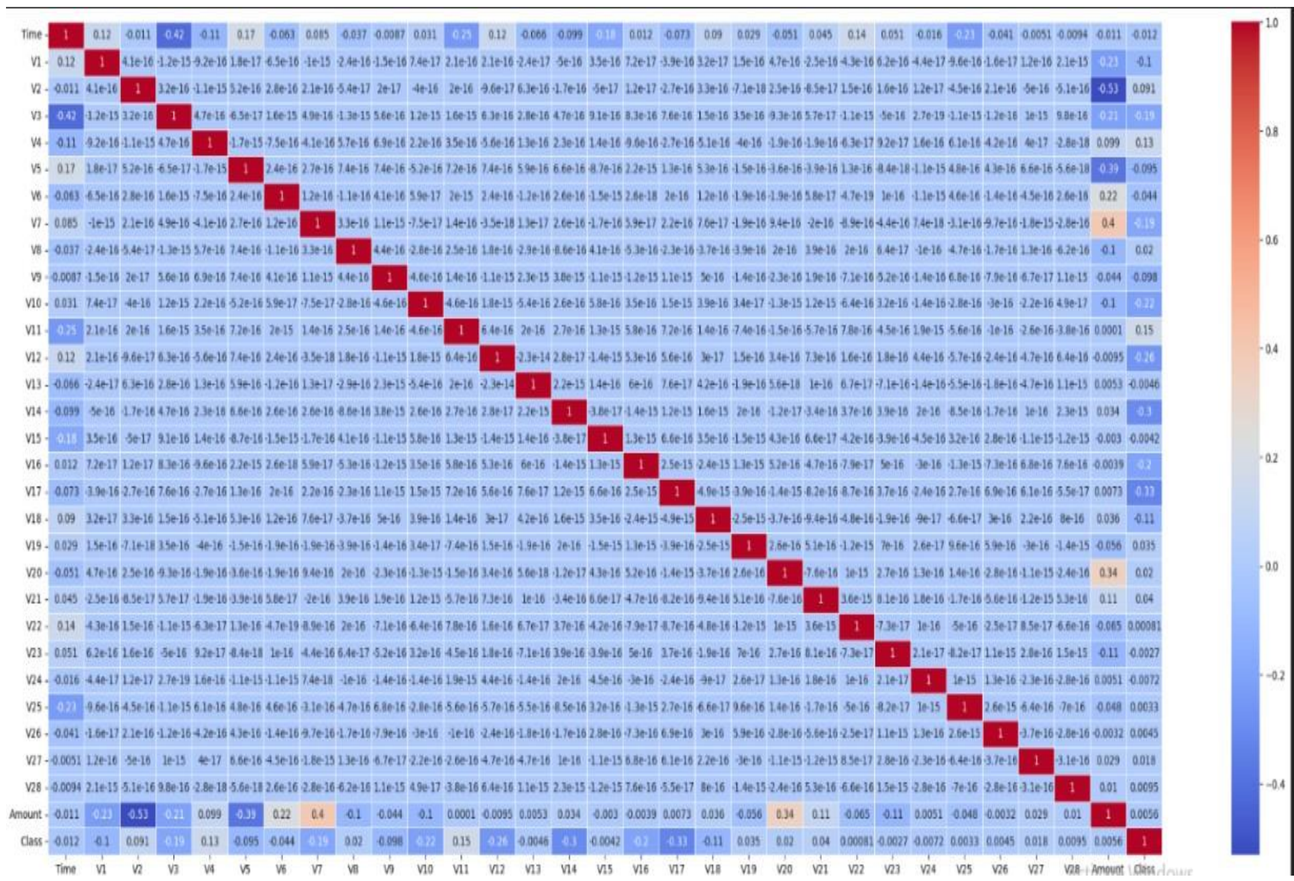
Normalization

- Normalization using StandardScaler is applied to 'Amount', 'Hour', 'Minute', and 'Second' features in the dataset.
- 'Amount' normalization ensures its proportional influence on learning relative to other features due to widely varying transaction amounts.

Data Visualization

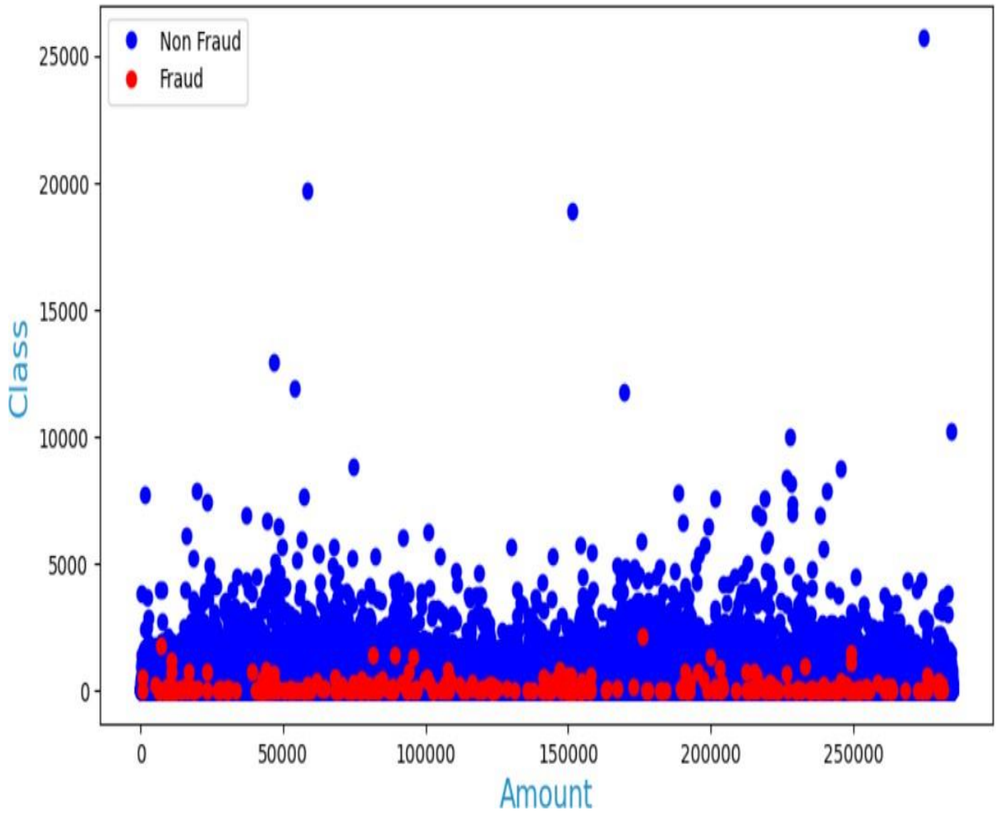


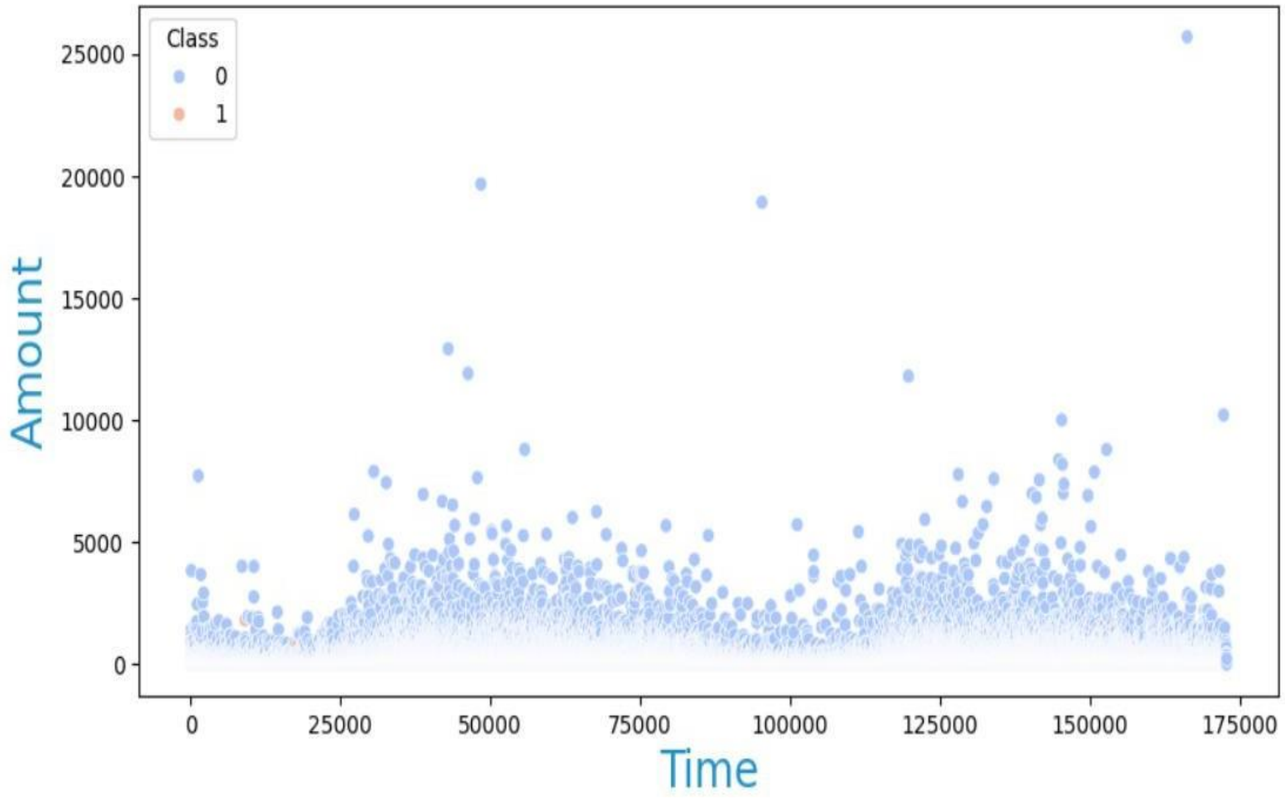
The correlation coefficient gauges the strength and direction of relationships between variables in a dataset, ranging from -1 to 1.



The heat map indicates no significant high correlation values among predictor columns, and none exhibit a strong correlation with the Class column.

Data Visualization

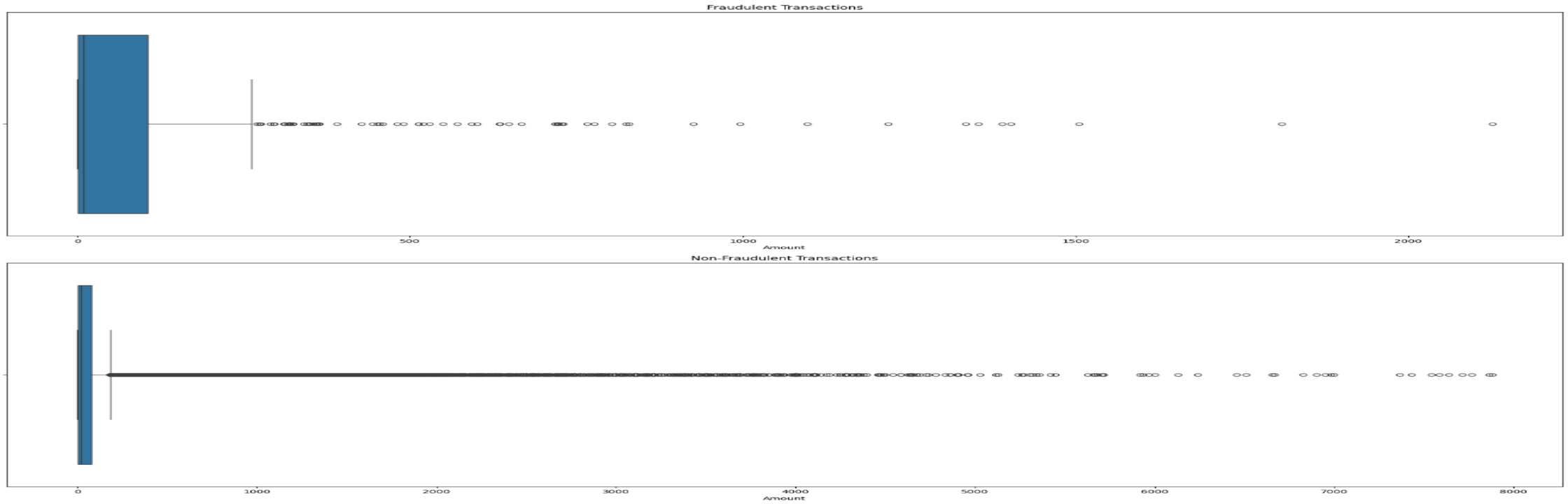




The scatter plot visualization indicates that there is no clear correlation between the transaction amount ('Amount') and the occurrence of fraud ('Class').

The scatter plot displays transactions with time on the x-axis, amount on the y-axis, and distinguishes fraud (pink) from non-fraud (blue), aiding in identifying patterns and outliers related to fraud.

Data Visualization



Fraudulent transactions are frequent at lower amounts, indicated by a prominent blue bar near the 0 mark, while non-fraudulent transactions display a more even distribution across various amounts, with a line extending towards higher values.

Methodology

- Undersampling Technique: It balances class distribution by removing instances from the majority class and utilizes the RandomUnderSampler from the imblearn library.

- Oversampling Technique: This method balances class distribution by duplicating instances from the minority class and leverages the RandomOverSampler from the imblearn library.

- We have utilized 10 different models :
    - Logistic Regression
    - Naive Bayes
    - AdaBoost
    - LightGBM
    - CatBoost
    - Artificial Neural Network (ANN)
    - KNN Algorithm
    - Random Forest Algorithm
    - XGBoost Algorithm
    - Decision Tree Algorithm

Model Evaluations and Comparison

Train Dataset

Model	Sampling Method	Recall	Precision	F1 Score	Accuracy	AUC Score
Logistic Regression	Undersampling	0.9399	0.9863	0.9626	0.9634	0.9634
Logistic Regression	Oversampling	0.9354	0.9746	0.9546	0.9555	0.9555
Naive Bayes	Undersampling	0.8512	0.9645	0.9043	0.9099	0.9099
Naive Bayes	Oversampling	0.8676	0.9702	0.9160	0.9205	0.9205
AdaBoost	Undersampling	0.9974	1.0000	0.9987	0.9987	0.9987
AdaBoost	Oversampling	0.9689	0.9837	0.9762	0.9764	0.9764
LightGBM	Undersampling	1.0000	1.0000	1.0000	1.0000	1.0000
LightGBM	Oversampling	1.0000	1.0000	1.0000	1.0000	1.0000
CatBoost	Undersampling	1.0000	1.0000	1.0000	1.0000	1.0000
CatBoost	Oversampling	1.0000	0.9998	0.9999	0.9999	0.9999
ANN	Undersampling	1.0000	1.0000	1.0000	1.0000	1.0000
ANN	Oversampling	1.0000	1.0000	1.0000	1.0000	1.0000
XGBoost	Undersampling	0.8889	0.0315	0.0608	0.9564	0.9227
XGBoost	Oversampling	1.0000	1.0000	1.0000	1.0000	1.0000
Decision Tree	Undersampling	1.0000	1.0000	1.0000	1.0000	1.0000
Decision Tree	Oversampling	1.0000	1.0000	1.0000	1.0000	1.0000
KNN	Undersampling	0.9034	0.9886	0.9441	0.9465	0.9465
KNN	Oversampling	1.0000	0.9996	0.9998	0.9998	0.9998
Random Forest	Undersampling	0.9869	1.0000	0.9934	0.9935	0.9935
Random Forest	Oversampling	1.0000	1.0000	1.0000	1.0000	1.0000

Model Evaluation and Comparison

Test Dataset

Model	Sampling Method	Recall	Precision	F1 Score	Accuracy	AUC Score
Logistic Regression	Undersampling	0.8889	0.0325	0.0627	0.9578	0.9234
Logistic Regression	Oversampling	0.9000	0.0392	0.0752	0.9649	0.9325
Naive Bayes	Undersampling	0.7889	0.0340	0.0652	0.9641	0.8766
Naive Bayes	Oversampling	0.8111	0.0444	0.0842	0.9720	0.8917
AdaBoost	Undersampling	0.9111	0.0153	0.0302	0.9070	0.9091
AdaBoost	Oversampling	0.8556	0.0741	0.1364	0.9828	0.9193
LightGBM	Undersampling	0.9111	0.0297	0.0576	0.9527	0.9319
LightGBM	Oversampling	0.8000	0.8090	0.8045	0.9994	0.8998
CatBoost	Undersampling	0.8778	0.0447	0.0850	0.9700	0.9240
CatBoost	Oversampling	0.7667	0.8118	0.7886	0.9993	0.8832
ANN	Undersampling	0.9000	0.0310	0.0599	0.9552	0.9277
ANN	Oversampling	0.7444	0.6204	0.6768	0.9989	0.8719
XGBoost	Undersampling	0.8889	0.0315	0.0608	0.9564	0.9227
XGBoost	Oversampling	0.9111	0.0156	0.0306	0.9084	0.9097
Decision Tree	Undersampling	0.9111	0.0156	0.0306	0.9084	0.9097
Decision Tree	Oversampling	0.6556	0.1565	0.2527	0.9938	0.8250
KNN	Undersampling	0.8444	0.0619	0.1154	0.9795	0.9121
KNN	Oversampling	0.7889	0.6514	0.7136	0.9990	0.8941
Random Forest	Undersampling	0.8889	0.0341	0.0656	0.9599	0.9244
Random Forest	Oversampling	0.7222	0.9286	0.8125	0.9995	0.8611

After comparing between models, Logistic Regression seems to be best model

Future plan of Action

Integrating More Diverse Data Sources

Include behavioral data, geolocation data , social media and public record

Implementing Advance Techniques

Implementing advanced techniques for credit card fraud detection such as : Deep learning Models, Graph-Based Methods and Ensemble Learning Techniques

.

Deploying the Model

Depolying our best model using different technologies such as : AWS Sagemaker, Flask, etc

Conclusion

- **After comparing between models, the results were found as follows:**
  Best model: Logistic Regression
  Training Set Accuracy score = 95.55%.
  Test Set Accuracy score = 96.49%.

- **Our project underscores the significance of interdisciplinary collaboration in understanding fraud patterns and developing effective detection strategies.**

- **We've recognized the importance of continuous monitoring, ethical considerations, and data quality for robust fraud detection systems**

References

[1] Anonymous. (n.d.). Credit Card Fraud Detection Predictive Modeling. Retrieved from https://library.ndsu.edu/ir/bitstream/handle/10365/31611/Credit%20Card%20Fraud%20Detection%20Predictive%20Modeling.pdf?sequence=1&isAllowed=y

[2] National Center for Biotechnology Information. (n.d.). Detecting Fraudulent Transactions. PMC. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10917329/#:~:text=Detecting%20fraudulent%20transactions%20is%20challenging,generate%20unrealistic%20or%20overgeneralized%20samples.

[3] Alturaby, N. (n.d.). Credit Card Fraud Detection: Risks and Challenges. Medium. Retrieved from https://medium.com/@nuhaaltoraby91/credit-card-fraud-detection-risks-and-challenges-f5e94796e1f1

Any Queries??

Thank You!!