U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Credit Card Fraud Detection

Fraud Detection (CC4036) – Final Project

Made By: Leonardo Freitas – up202400832

Problem Definition

What is the problem?

Credit card fraud detection aims to identify fraudulent transactions in real-time. With the rise in online transactions, detecting and preventing fraud has become a critical task for financial institutions.This project focuses on using machine learning to identify fraudulent transactions with high precision

What is the goal?

The primary goal of the project is to develop a machine learning-based system that can accurately detect fraudulent credit card transactions in real-time. The focus is on achieving high precision and recall to minimize false positives (flagging legitimate transactions as fraudulent) and false negatives (failing to detect actual fraud).

What is the proposed solution?

The proposed solution is to build and deploy a predictive model using machine learning techniques. By leveraging features from credit card transaction data, merchant details, customer demographics, and geographic data, the model can classify transactions as fraudulent or non-fraudulent. The models used include Logistic Regression, Support Vector Classifier, Decision Tree, and Random Forest.

How can it be achieved?

How can we evaluate/improve results gained from the project?

Data Understanding

Transactions.csv: Individual transaction records with fields like transaction time, credit card number, transaction amount, and fraud label.

Merchants.csv: Details about merchants, including category, location, and merchant ID.

Customers.csv: Customer demographic information such as gender, address, and job details.

Cities.csv: Geographic data on city names, population, and state information.

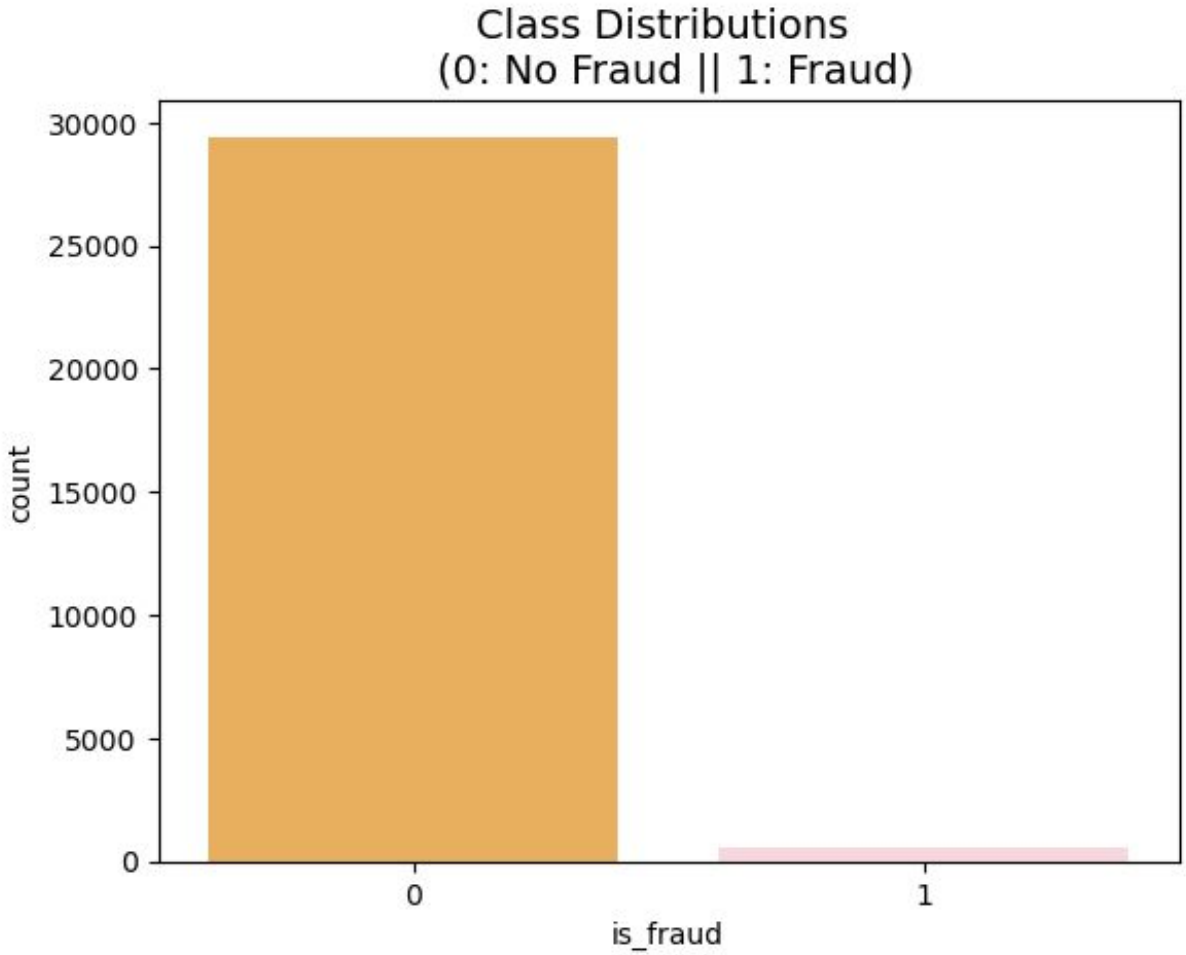
Note: All this data was merged into a unique dataset to create a unified data source.

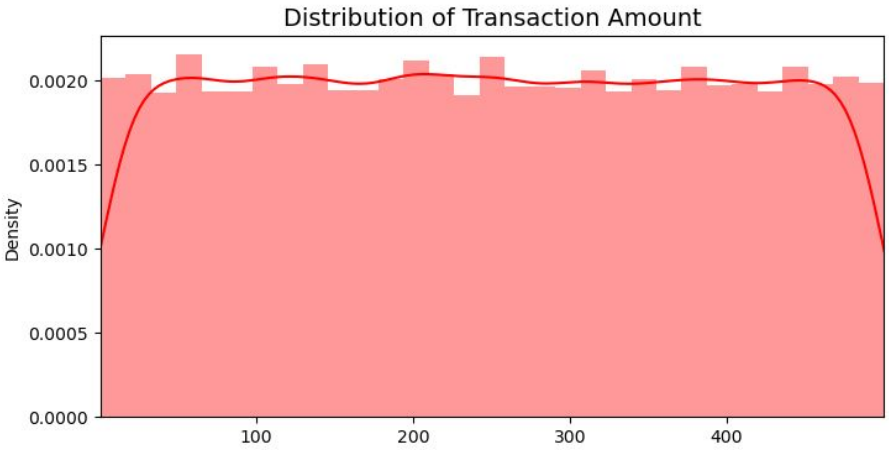
Key Insights:

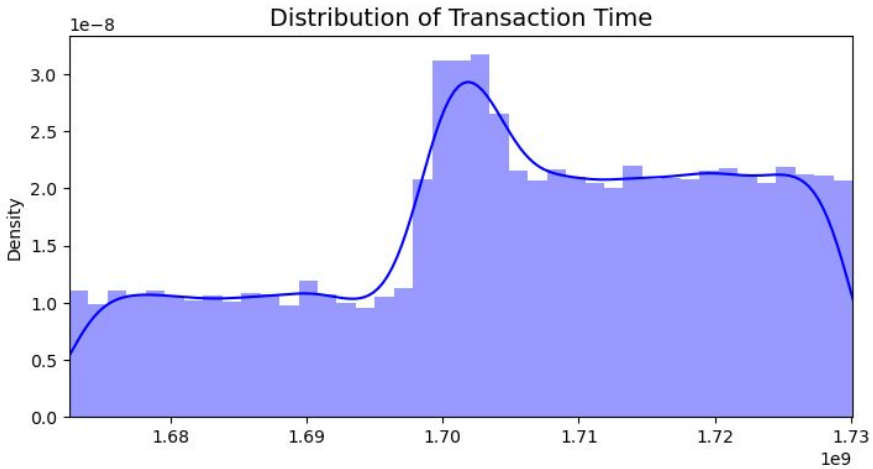
Data shows that fraudulent transactions are a small percentage of total transactions, posing a class imbalance challenge.

Merchant and customer data provide additional features for model training.

Geographical information (latitude, longitude) from merchants and cities offers opportunities for location-based analysis.


Class Distributions
(0: No Fraud || 1: Fraud)
30000
25000
20000
15000
10000
5000
0
count
0
1
is_fraud


Distribution of Transaction Amount
Density
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400


Distribution of Transaction Time
1e−8
Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.68
1.69
1.70
1.71
1.72
1.73
1e9

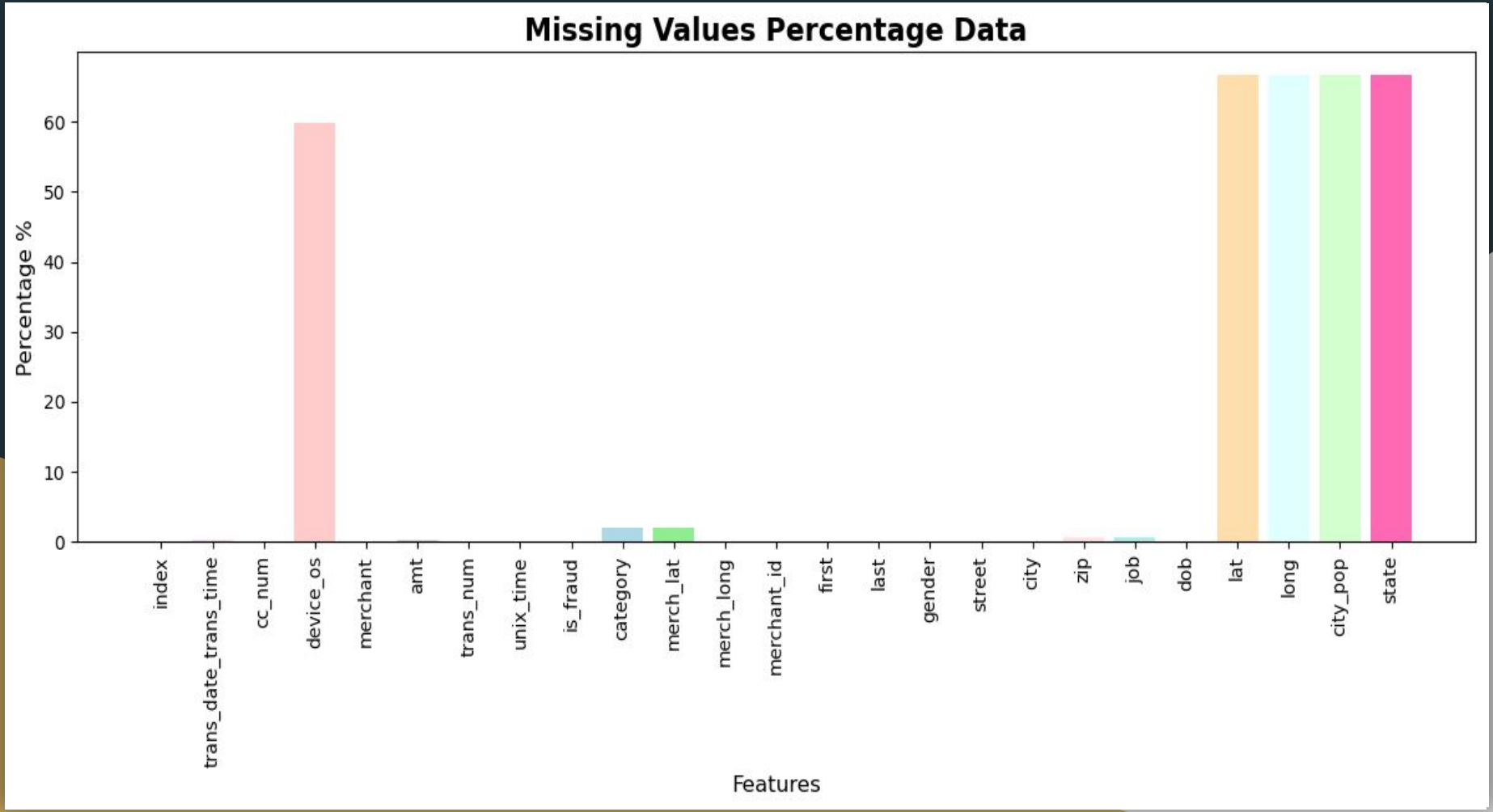
Data Preparation - Initial

Handled missing data through imputation or removal.

Dropped unwanted Columns

Feature Engineering

Scaler


Missing Values Percentage Data
Percentage %
0
10
20
30
40
50
60
index
trans_date_trans_time
cc_num
device_os
merchant
amt
trans_num
unix_time
is_fraud
category
merch_lat
merch_long
merchant_id
first
last
gender
street
city
zip
job
dob
lat
long
city_pop
state
Features

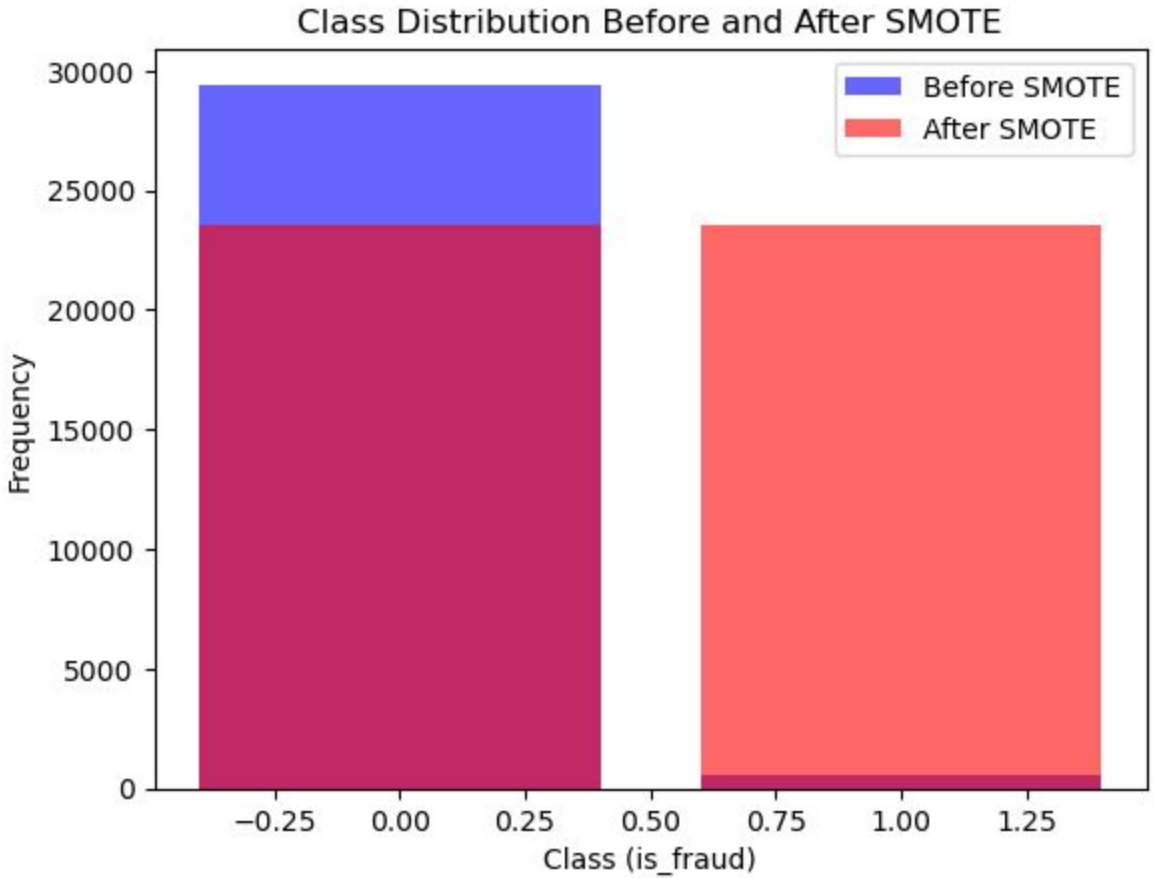
Predictive Modeling

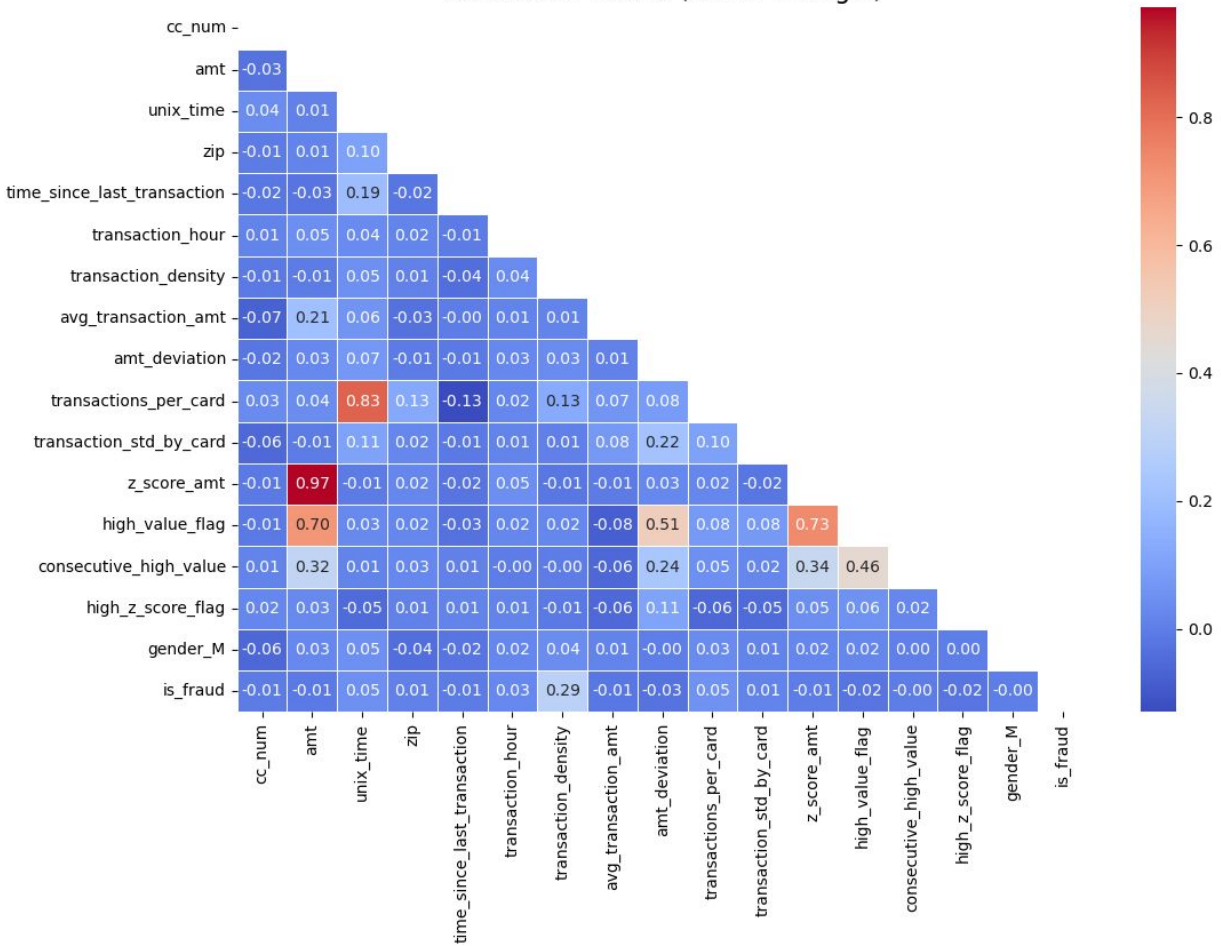
Splitting the Data into Train and Test Set using StratifiedKFold

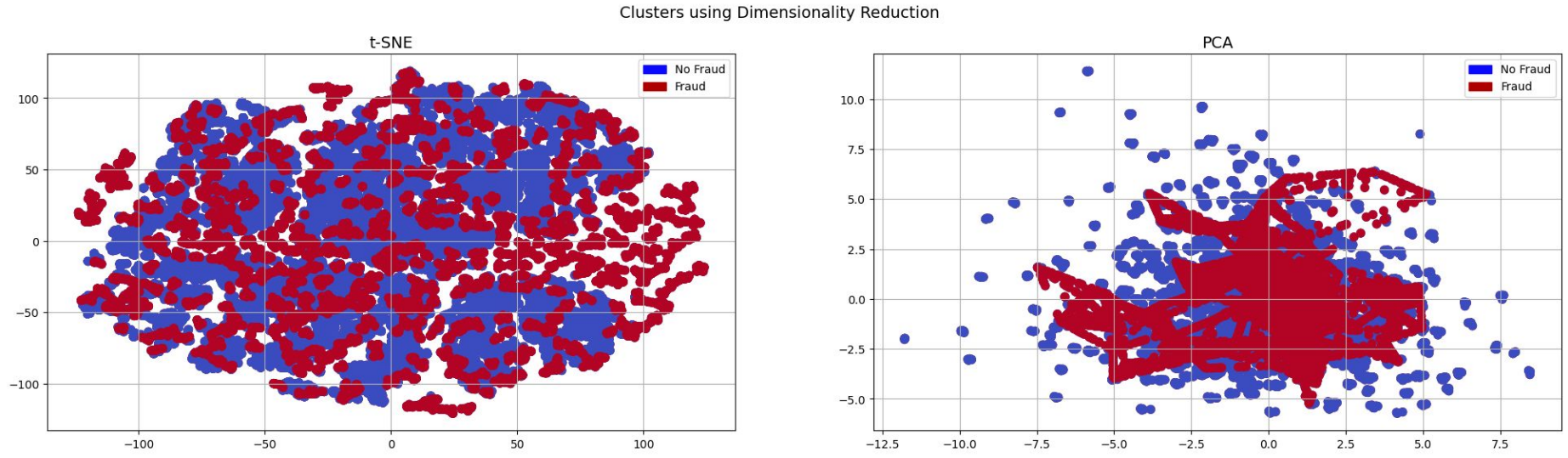
Applying SMOTE to balance the fraud and non-fraud class.

Correlation Matrix between features

Dimensionality Reduction and Clustering


Class Distribution Before and After SMOTE
Before SMOTE
After SMOTE
Frequency
30000
25000
20000
15000
10000
5000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Class (is_fraud)


Correlation Matrix (Lower Triangle)
cc_num
amt -0.03
unix_time 0.04 0.01
zip -0.01 0.01 0.10
time_since_last_transaction -0.02 -0.03 0.19 -0.02
transaction_hour 0.01 0.05 0.04 0.02 -0.01
transaction_density -0.01 -0.01 0.05 0.01 -0.04 0.04
avg_transaction_amt -0.07 0.21 0.06 -0.03 -0.00 0.01 0.01
amt_deviation -0.02 0.03 0.07 -0.01 -0.01 0.03 0.03 0.01
transactions_per_card 0.03 0.04 0.83 0.13 -0.13 0.02 0.13 0.07 0.08
transaction_std_by_card -0.06 -0.01 0.11 0.02 -0.01 0.01 0.01 0.08 0.22 0.10
z_score_amt -0.01 0.97 -0.01 0.02 -0.02 0.05 -0.01 -0.01 0.03 0.02 -0.02
high_value_flag -0.01 0.70 0.03 0.02 -0.03 0.02 0.02 -0.08 0.51 0.08 0.08 0.73
consecutive_high_value 0.01 0.32 0.01 0.03 0.01 -0.00 -0.00 -0.06 0.24 0.05 0.02 0.34 0.46
high_z_score_flag 0.02 0.03 -0.05 0.01 0.01 0.01 -0.01 -0.06 0.11 -0.06 -0.05 0.05 0.06 0.02
gender_M -0.06 0.03 0.05 -0.04 -0.02 0.02 0.04 0.01 -0.00 0.03 0.01 0.02 0.02 0.00 0.00
is_fraud -0.01 -0.01 0.05 0.01 -0.01 0.03 0.29 -0.01 -0.03 0.05 0.01 -0.01 -0.02 -0.00 -0.02 -0.00
cc_num
amt
unix_time
zip
time_since_last_transaction
transaction_hour
transaction_density
avg_transaction_amt
amt_deviation
transactions_per_card
transaction_std_by_card
z_score_amt
high_value_flag
consecutive_high_value
high_z_score_flag
gender_M
is_fraud
0.8
0.6
0.4
0.2
0.0


Clusters using Dimensionality Reduction
t-SNE
No Fraud
Fraud
100
50
0
−50
−100
−100
−50
0
50
100
PCA
No Fraud
Fraud
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

Predictive Modeling - MODELS

For this project, and especially after seeing how the data is distributed, 4 models were considered:

- **Logistic Regression:** Supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not.

- **Support Vector Classifier:** SVC is effective for datasets where the decision boundary is non-linear. It finds the optimal hyperplane that maximizes the margin between classes.

- **Decision Tree:** Decision Trees are intuitive, interpretable, and capable of capturing non-linear patterns in the data.

- **Random Forest:** Random Forest combines multiple Decision Trees to create a more stable, robust, and accurate model.

Predictive Modeling - MODELS Improvements

- GridSearchCV used to find optimal hyperparameters for each model.

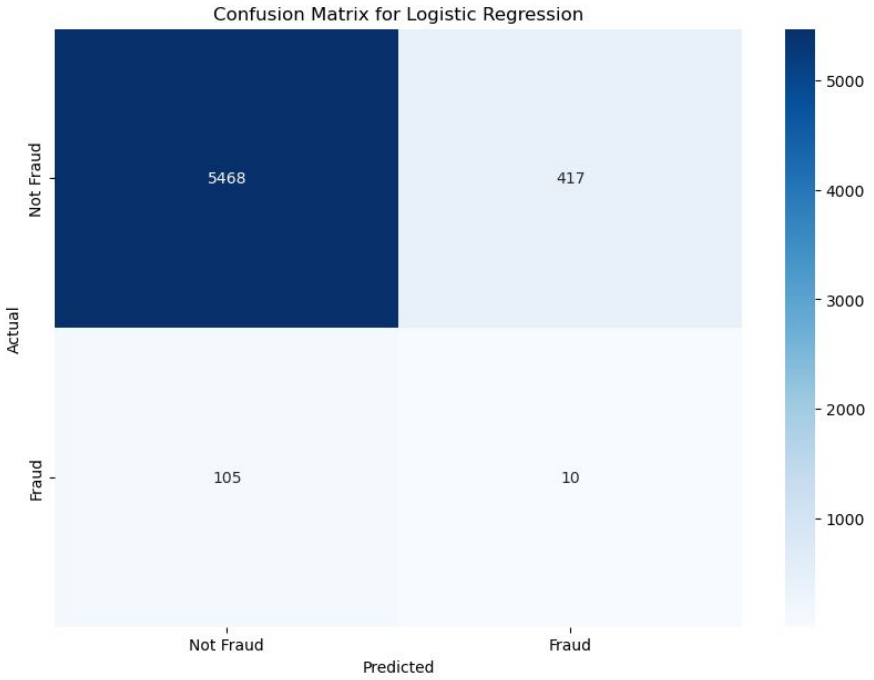
- Cross-validation ensures that models generalize well to unseen data.

Predictive Modeling - **Logistic Regression**

Model: Logistic Regression
Classification Report:

	precision	recall	f1-score	support
0	0.98	0.93	0.95	5885
1	0.02	0.09	0.04	115
accuracy			0.91	6000
macro avg	0.50	0.51	0.50	6000
weighted avg	0.96	0.91	0.94	6000

AUC for Logistic Regression: 0.57

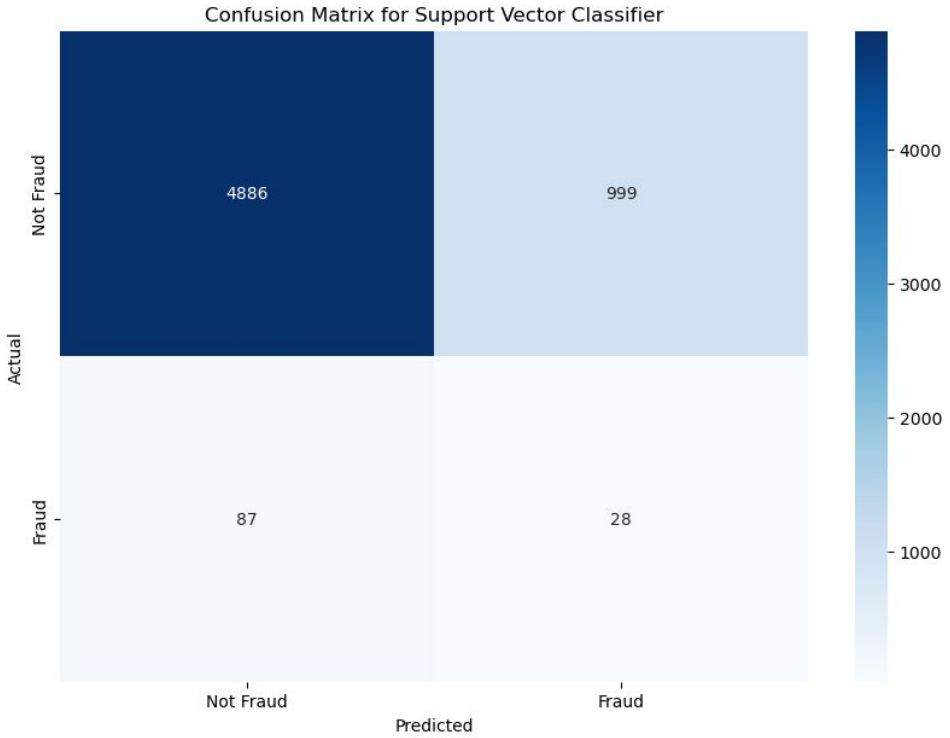


Predictive Modeling - **Support Vector Classifier:**

Model: Support Vector Classifier
Classification Report:

	precision	recall	f1-score	support
0	0.98	0.83	0.90	5885
1	0.03	0.24	0.05	115
accuracy			0.82	6000
macro avg	0.50	0.54	0.47	6000
weighted avg	0.96	0.82	0.88	6000

AUC for Support Vector Classifier: 0.57



Predictive Modeling - **Decision Tree:**

Model: Decision Tree
Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5885
1	1.00	0.03	0.05	115
accuracy			0.98	6000
macro avg	0.99	0.51	0.52	6000
weighted avg	0.98	0.98	0.97	6000

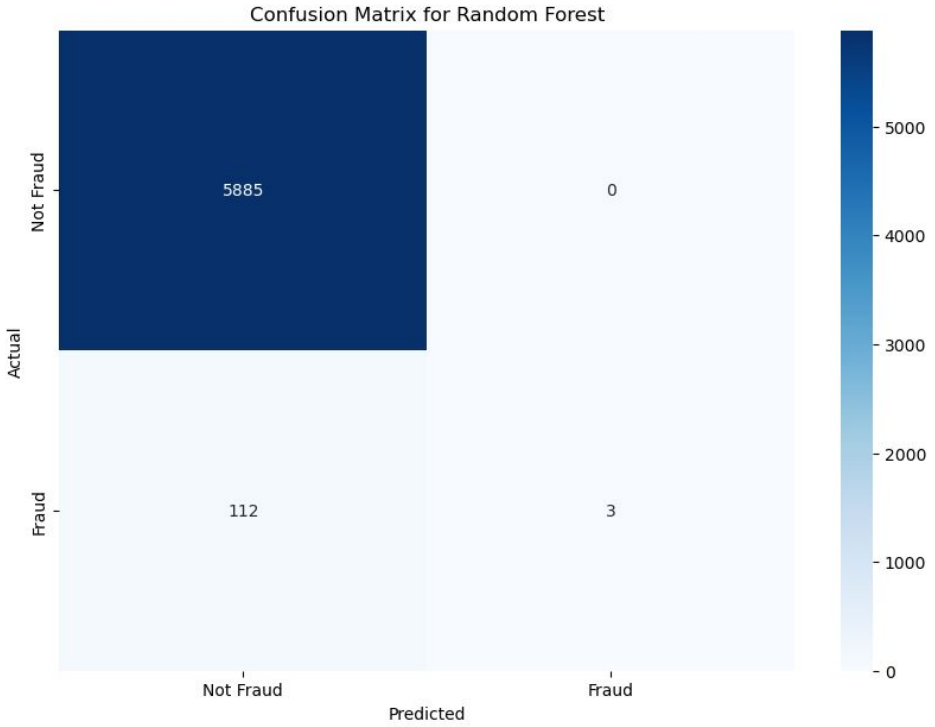
AUC for Decision Tree: 0.53

Predictive Modeling - **Random Forest:**

Model: Random Forest
Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5885
1	1.00	0.03	0.05	115
accuracy			0.98	6000
macro avg	0.99	0.51	0.52	6000
weighted avg	0.98	0.98	0.97	6000

AUC for Random Forest: 0.54

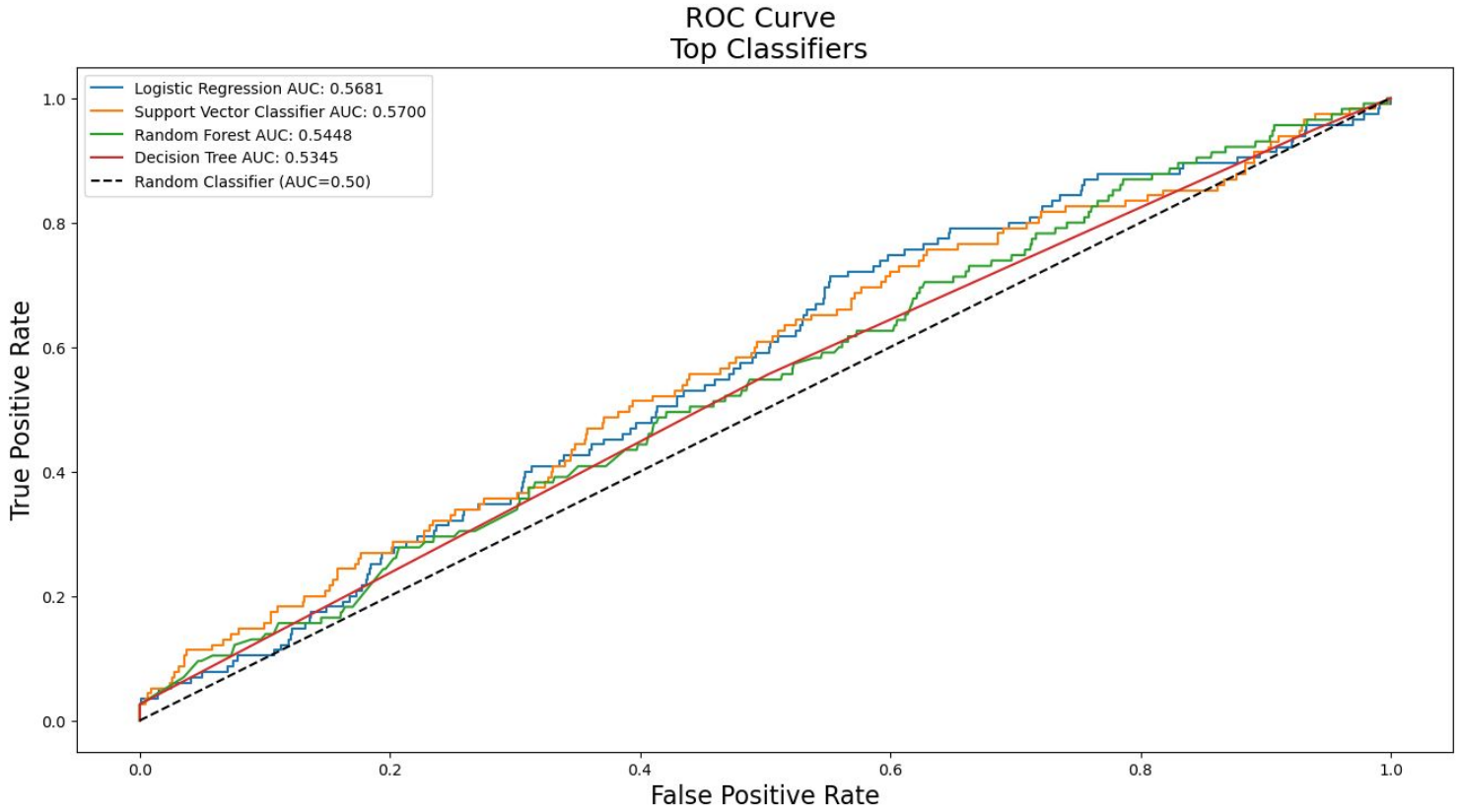


Predictive Modeling - Feature Importance

Feature Importance for Logistic Regression:

Feature	Importance
transaction_density	2.215188
z_score_amt	1.282910
unix_time	0.333065
consecutive_high_value	0.131773
transaction_hour	0.062315
avg_transaction_amt	0.057802
transaction_std_by_card	0.023737
zip	0.011669
cc_num	-0.045907
time_since_last_transaction	-0.065351
gender_M	-0.070805
amt_deviation	-0.103181
high_value_flag	-0.234292
transactions_per_card	-0.326949
high_z_score_flag	-0.978962
amt	-1.201089

Predictive Modeling - ROC Curve



Future Work & Limitations

Improve Feature Engineering

Test New Parameters

Test New Models

Conclusion

The Credit Card Fraud Detection project successfully addressed the critical issue of identifying fraudulent transactions in real-time. Leveraging machine learning models, the project demonstrated how data-driven approaches can enhance fraud detection with higher precision and recall. Despite the success, the project faced some limitations. The class imbalance posed a challenge, but it was mitigated using SMOTE. Another limitation was the computational cost and complexity of some models, particularly SVC. The project achieved a strong balance between precision and recall. Moving forward, continued work on feature engineering, scalability, and real-time deployment will ensure even greater impact.