Credit Card Fraud Prediction

Report by

SABIULLAH N

Batch No: DW67

Table Of Contents

1. Introduction....4
2. Data Overview....4
   Exploratory Data Analysis (EDA)....4
3. Feature Selection....4
   Top Variables for Prediction....4
   PCA Implementation....4
4. Modelling....5
   Under Sampling :....5
   Over Sampling:....6
5. Conclusion....7
   Under Sampling....7
   Oversampling....7
   Model Selection:....7

Table Of Figure

Figure Name	**Page No**
Fig:1 – Correlation heatmap of Top Features with Class	5
Fig:2 – Performance metrics comparison among all the models (Under Sampling)	6
Fig:3 – Performance metrics comparison among all the models (Over Sampling)	6

1. Introduction

The project aimed to predict credit card fraud using machine learning techniques. The dataset, sourced from Kaggle, consisted of credit card transactions, highly imbalanced, with 492 frauds among 284,807 transactions. The report explores feature selection, modelling using various classifiers, and evaluates model performance under different sampling techniques.

2. Data Overview

Exploratory Data Analysis (EDA)

The data set contains 284807 rows and 31 columns, and found 1081 duplicated rows, after removing all the duplicated rows found the actual data that 473 fraud transactions and 283253 normal transactions.

3. Feature Selection

Top Variables for Prediction

Explanation of the most influential variables for predicting fraud, potentially derived from correlation analysis or feature importance.

PCA Implementation

By doing correlation matrix heatmap found top features of the class and keep the high featured columns alone with Class and removed rest of the columns from the data set to proceed further modelling.

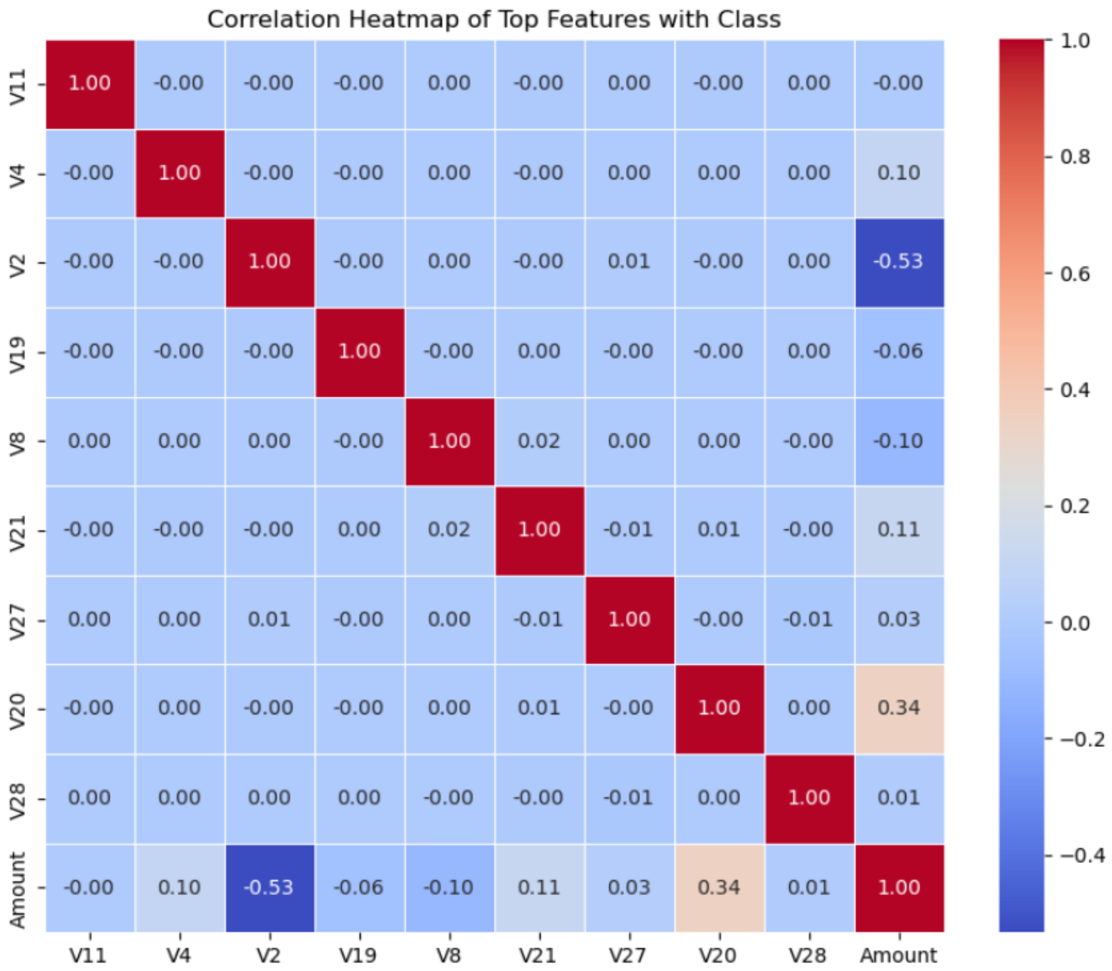

Fig:1 – Correlation heatmap of Top Features with Class

4. Modelling

Imbalanced data set is balanced by applying Under Sampling & Over Sampling method to get which of the balanced data set will give more accuracy.

Under Sampling :

	Logistic Regression	Decision Tree Classifier	SVM Method
Accuracy	0.915789	0.878947	0.905263
Precision	0.941176	0.867257	0.915094
Recall	0.905660	0.924528	0.915094
F1 Score	0.923077	0.894977	0.915094

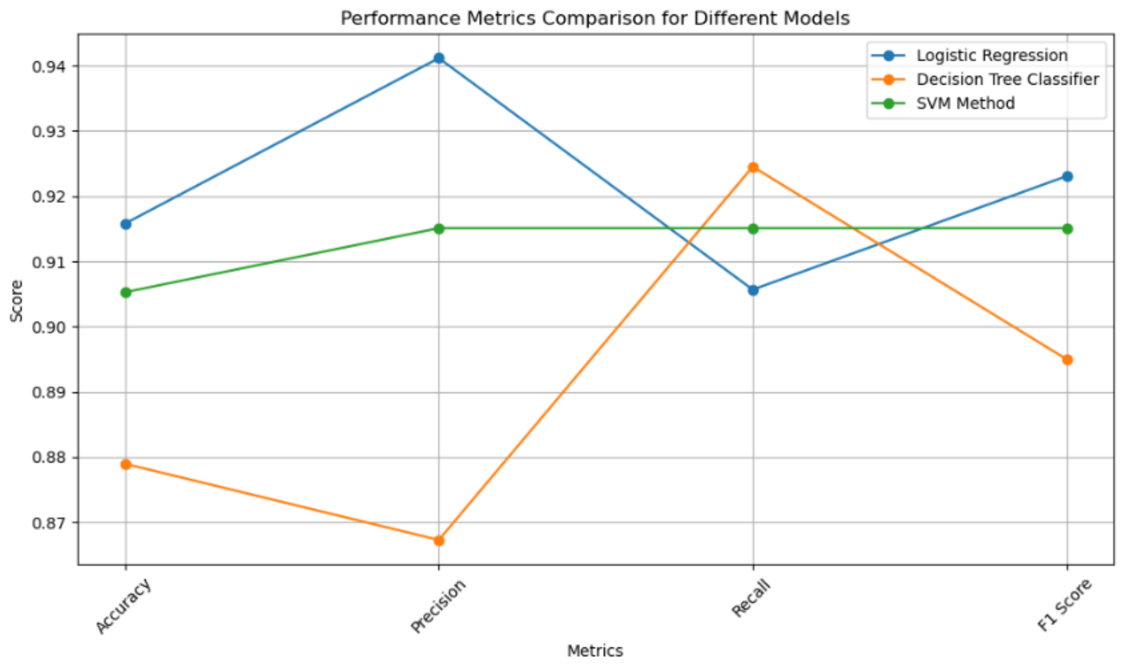

Fig:2 – Performance metrics comparison among all the models (Under Sampling)

Over Sampling:

	Logistic Regression	Decision Tree Classifier	SVM Method
Accuracy	0.916383	0.999762	0.946065
Precision	0.942884	0.999525	0.971922
Recall	0.887043	1.000000	0.919030
F1 Score	0.914112	0.999763	0.944736

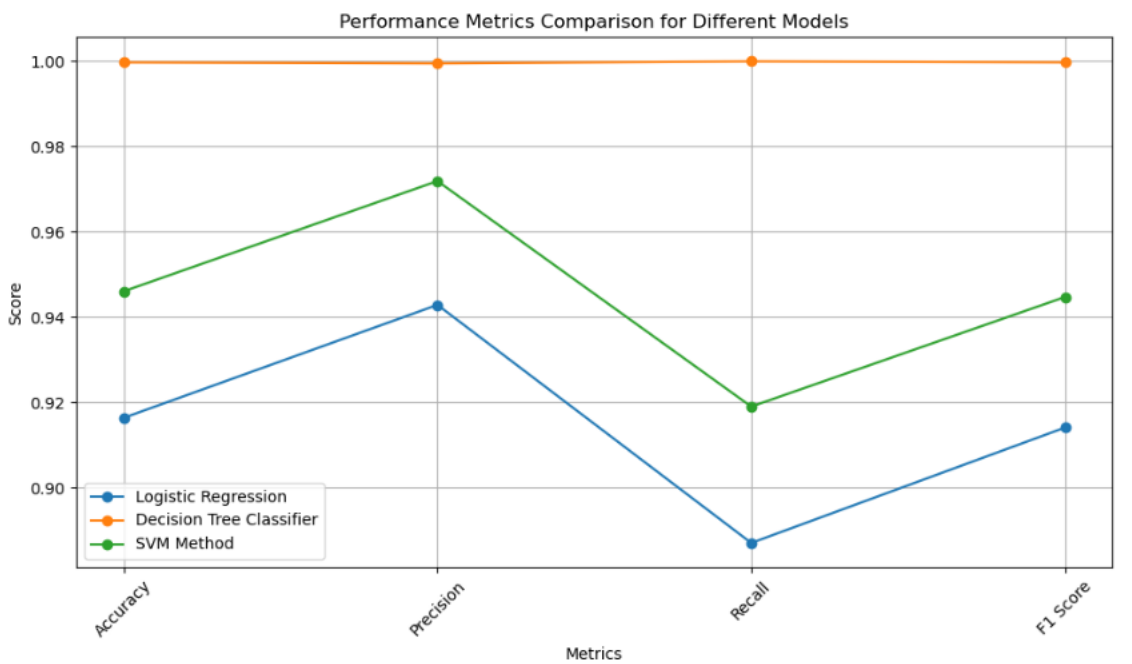

Fig:3 – Performance metrics comparison among all the models (Over Sampling)

5. Conclusion

Under Sampling

Logistic Regression: High accuracy and decent across other metrics.

Decision Tree Classifier: Good overall performance but slightly lower precision and F1 score compared to the other models.

SVM Method: High accuracy and balanced precision, recall, and F1 score.

Oversampling

Logistic Regression: Slightly improved performance across all metrics.

Decision Tree Classifier: Significant improvement in all metrics, almost perfect scores.

SVM Method: Improved accuracy and precision, slightly lower recall but an improved F1 score.

Model Selection

Based on the provided results, the oversampling method has notably improved the performance of all models. The Decision Tree Classifier exhibits impressive performance with almost perfect scores across the board in the oversampling scenario.