Tamilmani R
Data Science
GeeksInformation Internship

Fraud Detection in Credit Card Transactions

DATASET

The dataset used in this project was obtained from GeeksInformation. It comprises250000 records and 31 features. The data was collected for the purpose of fraud detection in credit card transactions. Fraudulent transactions can lead to financial losses, reputational damage, and customer dissatisfaction. To mitigate these risks, businesses employ various methods to detect and prevent fraud.

In this project, we aimed to develop an effective fraud detection system for credit card transactions. Two machine learning models, Logistic Regression and Random Forest, were employed to achieve this goal.

DATA OVERVIEW

The dataset consists of 250000 instances and 2 features. The features include numerical variables.

DATA TYPES AND FEATURES

Time:

Data Type: Numerical

Description: The time elapsed since the first transaction in seconds.

V1-V28:

Data Type: Numerical

Description: Anonymous feature resulting from a PCA transaction. They are numerical variable derived from the original data to protect user identities.

Amount:

Data Type: Numerical

Description: The transaction amounts.

Class:

Data Type: Numerical

Description: The target variables indicating whether the transaction is fraud (1) or not (0).

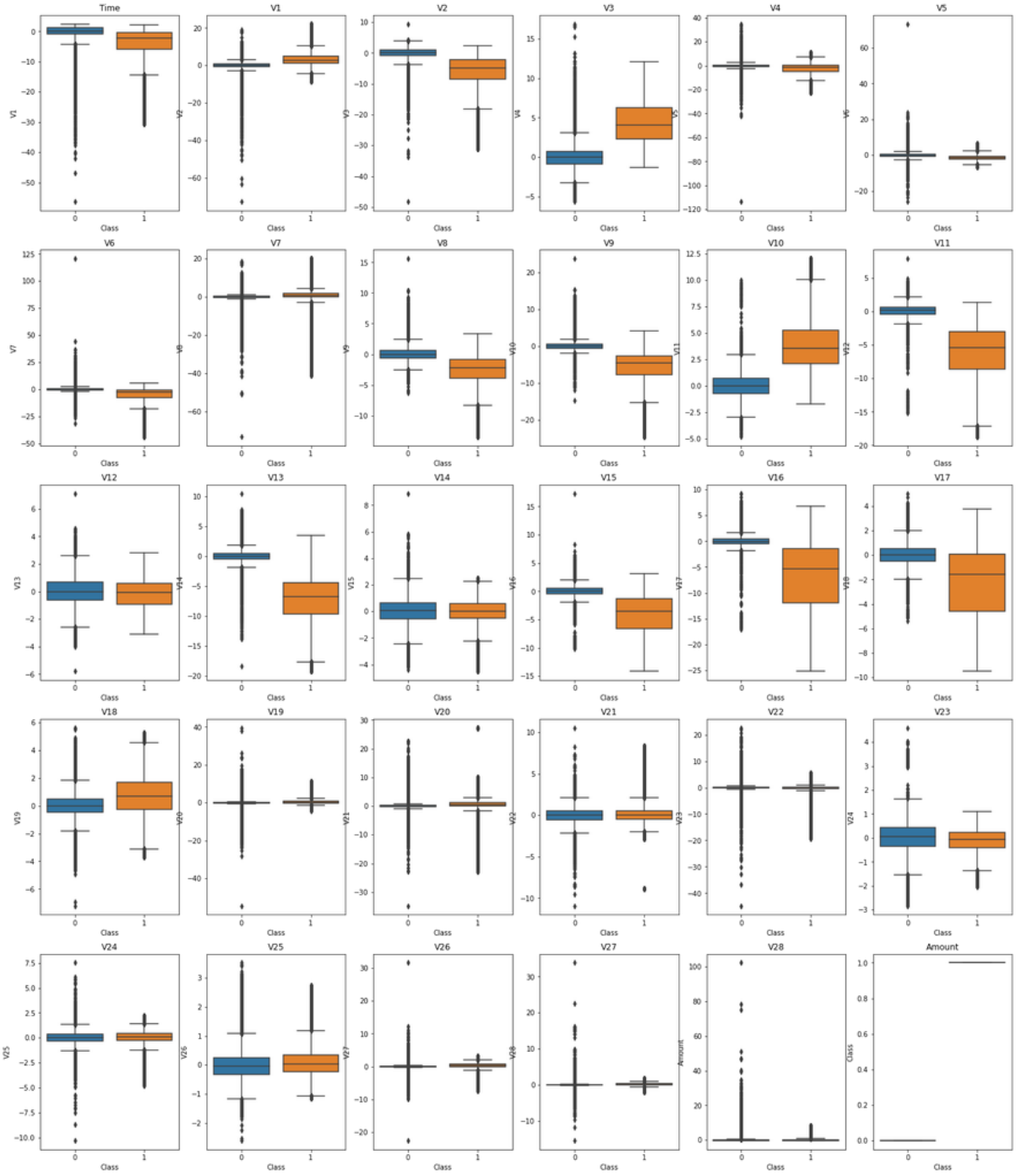
MISSING VALUES

The dataset is free of missing values, providing a clean and complete foundation for analysis.

DATA DESCRIPTION

Provide visualizations or summary statistics that illustrate the distribution of the data. This could include histograms, box plots, or other relevant charts.

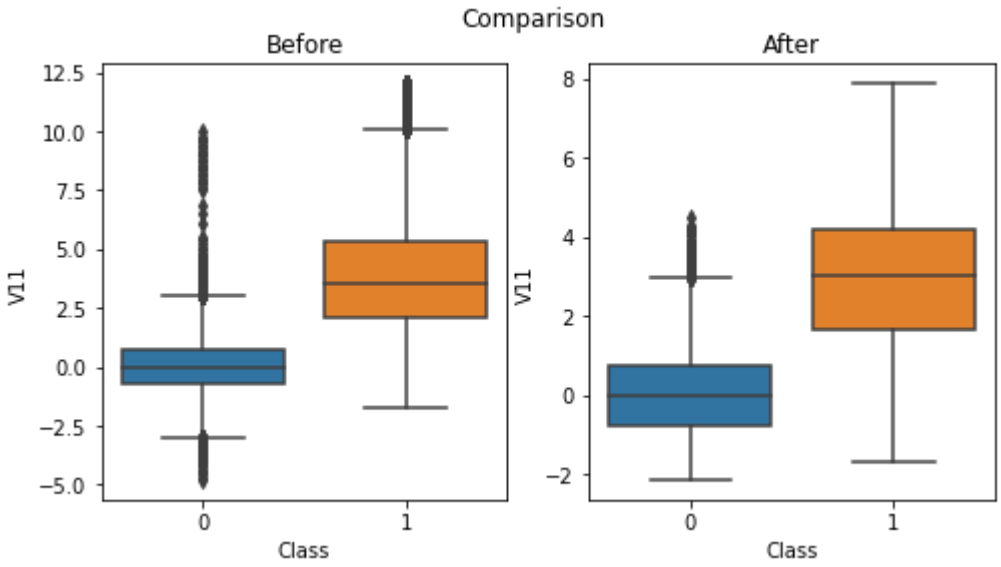
As you can observe from the plot, we have so many 0 (non-fraud) compared to 1 (fraud). This kind of imbalance in the target variable is known as class imbalance.



Here we can see that some of the feature there are a clear range between the class. We can also see that there are a lot of outliers. So we will remove the extreme outliers from the feature that have a high correlation with the class.

From the boxplot, we can see that the dataset that have a negative and positive correlation with the class and we can see that feature that have a high correlation with the class is V2 V3 V8 V10 V11 V13 V15 V16 V17 V18 so we will remove outliers from this dataset.

Here are the comparison of before and after removing the outliers on one of the feature.



The threshold used for removing the extreme outliers is 1.5 of IQR. The cut off value will be used to determine the range, which the lower range starts from the Q25 — cut off value until the upper range which is Q75 + cut off value.

The value outside of those range will be removed.

After removing outliers, we can see that there is a slight decreasing in the fraud case (Class 1) dataset. Although the data looks imbalanced, there is not a huge different to consider this dataset is imbalanced.

Lastly, we will do SMOTE, we can do it by using the imbalance libraries. In our dataset, we have binary classification problem where class A constitutes 80% of the sample and class B constitutes only 20%

We utilized the following python library ‘scikit-learn’ for machine learning functionalities and ‘imbalanced-learn’ for SMOTE implementation.

After we applying SMOTE, we have successfully balanced our imbalanced dataset

MODELS

In the section, we explore the application of machine learning model to address our research question. we employed Logistic Regression and Random Forest due to their suitability for our binary classification problem.

Logistic Regression:

Logistic regression is a linear model widely used for binary classification task. It calculates the probability of an belong to particular class and is particular useful when the relationship between the feature and target variable.
The Logistic Regression model was trained using the training dataset (**X_train** and **y_train**) after training, the model was used to make predictions on the test dataset (**X_test**). The predicted values.

Random Forest:

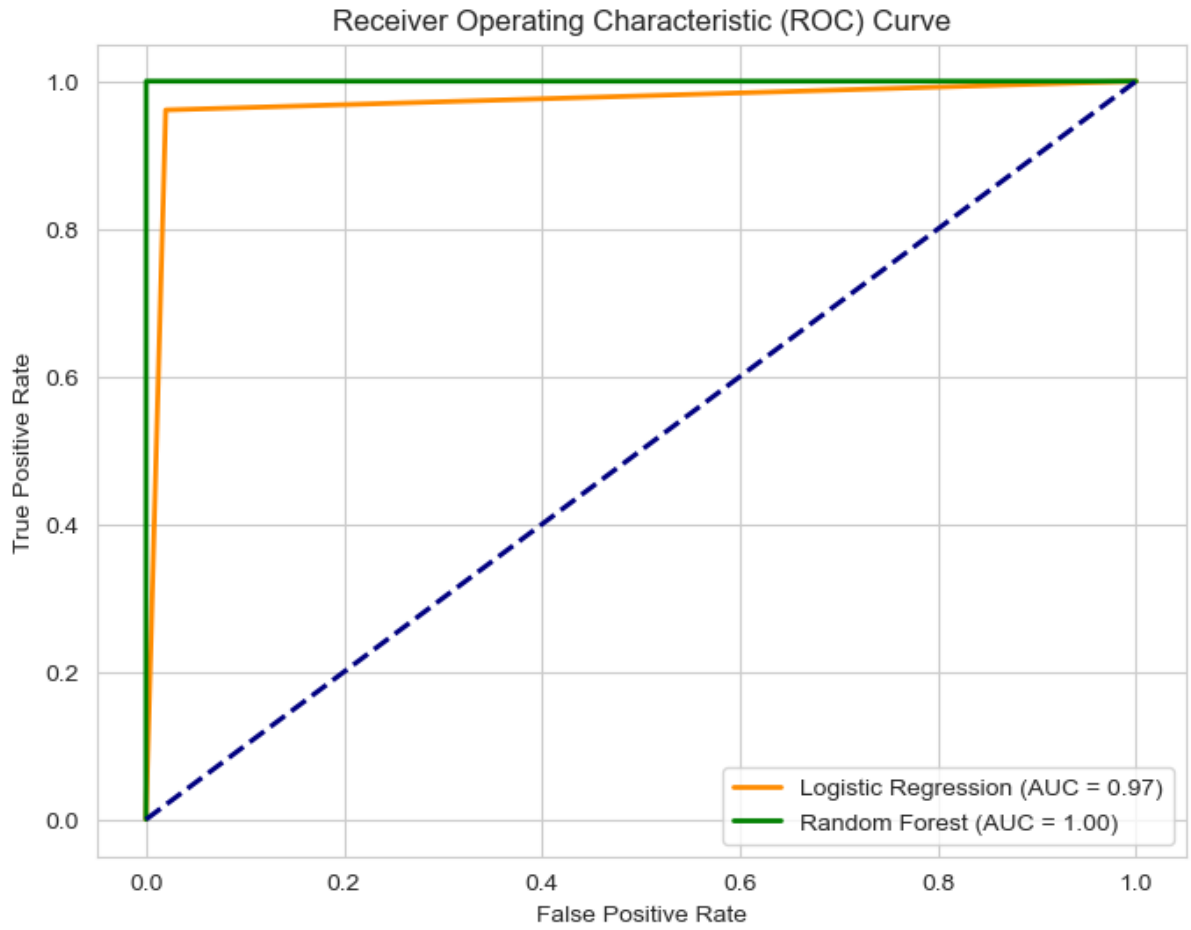
Random Forest is ensemble learning method that combine the prediction of multiple decision tree. First random forest was instantiated after model was trained using the training dataset (**X_train** and **y_train)** and the model was used to make prediction on the test data.

MODEL EVALUATION:

To assess the preformation of the Logistic regression and Random Forest model, we employed a range of evaluation metrics, including accuracy, Precision, recall and F1 score. The result is summarized in the following table.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	97%	97%	96%	97%
Random Forest	99%	99%	100%	99%

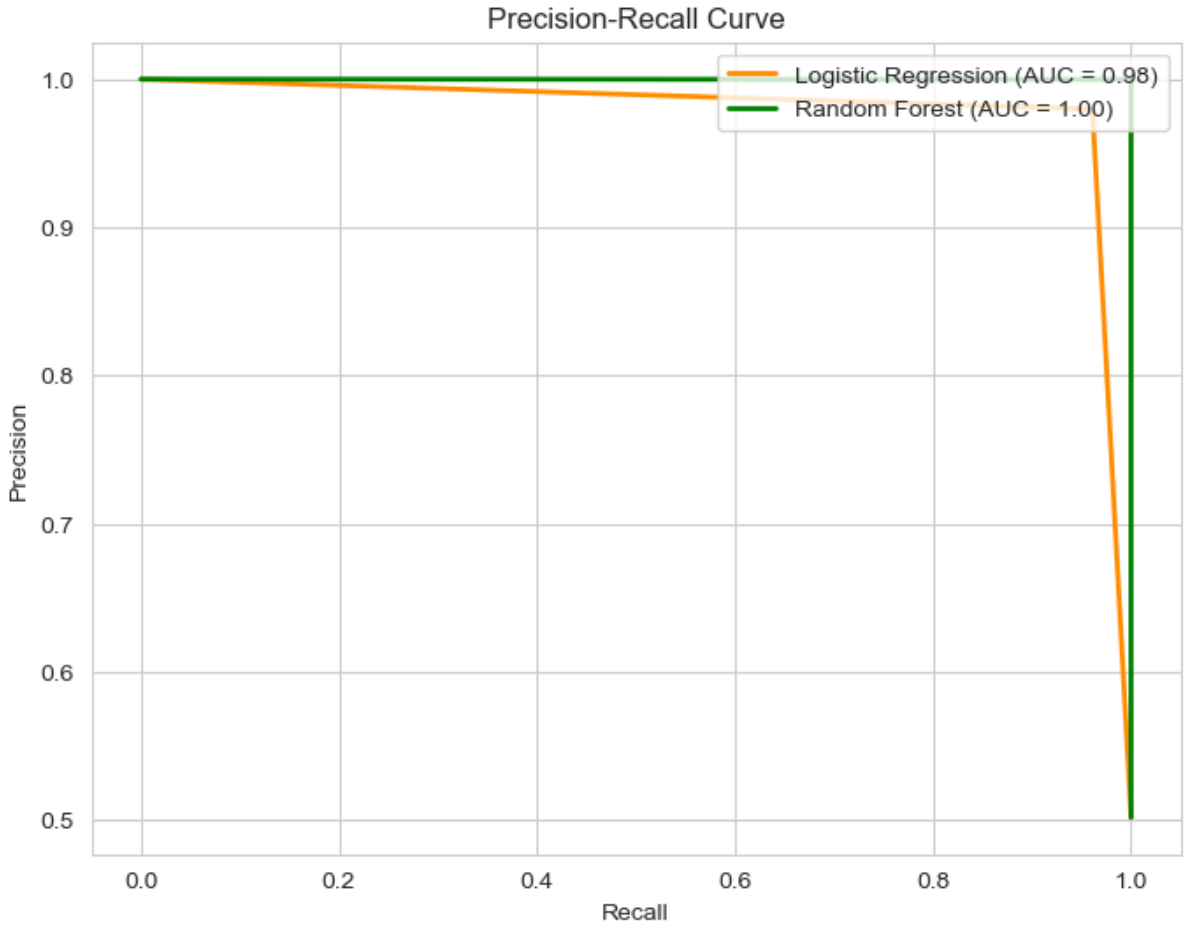
ROC CURVE



The ROC curve shows the performance of a binary classifier at different threshold values. The true positive rate (TPR) is the proportion of positive cases that are correctly identified by the classifier, and the false positive rate (FPR) is the proportion of

negative cases that are incorrectly identified as positive by the classifier. The ROC curve is plotted with the TPR on the y-axis and the FPR on the x-axis. A perfect classifier would have a TPR of 1 and an FPR of 0, meaning that it would correctly identify all positive cases and none of the negative cases.

A higher AUC indicates a better performing classifier. The AUC for the logistic regression model in the image is 0.97, and the AUC for the random forest model is 1.00. This means that the random forest model is performing slightly better than the logistic regression model on this dataset.



precision-recall curve (PRC) for two different machine learning models: logistic regression and random forest. Precision is the fraction of positive predictions that are actually positive. Recall is the fraction of positive cases that are correctly identified by the classifier.

The PRC is plotted with precision on the y-axis and recall on the x-axis. A perfect classifier would have a precision of 1 and a recall of 1, meaning that it would correctly identify all positive cases and none of the negative cases.

CONCLUSION

In conclusion, the Logistic Regression and Random Forest models have proven to be robust tools for credit card fraud detection, achieving accuracies of 97% and 99%, respectively. The project provides a solid foundation for the development of a reliable and efficient fraud detection system in the financial sector.