CREDIT CARD FRAUD DETECTION SYSTEM

SUBMITTED BY:

OLETI CHANDINI

MENTOR:

Muvendiran M

List of Contents :

1. Introduction

Credit card fraud detection is a critical application in financial services, as it helps to prevent unauthorized transactions and financial losses. The **Credit Card Fraud Detection Dataset** is used for building predictive models to classify whether a given transaction is fraudulent or legitimate. This dataset contains anonymized transaction details and includes both the features of the transactions and labels indicating whether a transaction is fraudulent (Class 1) or legitimate (Class 0).

This dataset is often used in machine learning tasks such as classification, anomaly detection, and fraud prediction. It can be used to train machine learning models that assist in detecting fraudulent activities based on the patterns learned from the data.

As this kind of problem is particularly challenging in a learning perspective, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumbered fraudulent ones and also the transaction pattern changes in statistical properties over the course of time.

Thus, in real world examples, the massive streams of payment requests are been quickly scanned by the automatic tools that determine which transaction to be authorized in order to prevent the performance of the fraud detection overtime As for many banks which has retained high profitable customers has been the number one business goal, these banking frauds poses a significant threat at different banks but in terms of substantial financial losses, trust and credibility has been a concerning issue for both banks and customers.

So in this project, we have detected the fraudulent credit card transactions with the help of machine learning models and analyze the customer data which has been collected.

Machine learning is a subfield of Artificial Intelligence (AI) which is generally used to understand the structure of data and fit the data into models which can be understood and used by people.

The algorithm instead allows the computer to train on data inputs and use statistical analysis for output values which falls in a specific range. So the machine learning makes easy for computers in building models from sample data in order to operate the decision making process based on the data inputs.

1.1 Methodology:

Most of this project has been done in google collab of language python because it is convenient for executing the program code and it does not take more time to execute with machine learning algorithms in order to detect with different activities. So, first of all, we obtained the dataset from Kaggle which is a data analysis website used for providing datasets. Inside the dataset, there are 31 columns of which there are 28 features represented as v1 to v28 in order to protect the sensitive data and the other column represents Time, Amount and Class.

The time shows the time gap between the first transaction, amount represents the total of money transacted and class represents the target value where 0 represents legitimate transaction and 1 represents fraudulent transactions. Since python libraries are free and open source, it has been built using Numpy and pandas for analyzing the data, matplotlib and seaborn to visualize the data and also provides simple and efficient tools for machine learning which has various classification and regression algorithms and it is designed to inter-operate the numerical libraries.

1.2 Exploratory Data Analysis (EDA) :

It is the process of investigating the dataset for discovering the patterns and forms hypotheses based on the understanding of the data set . It involves in generating the summary statistics for numerical data in the dataset and creates various graphical representations to understand the data.

1.3 Confusion Matrix :

It is a table which is used to describe the performance of a classification model or a classifier on a set of test data for which the true values are known. By visualizing the confusion matrix, an individual could determine the accuracy of the model by observing the diagonal values for measuring the number of accurate classifications. The structure of the matrix is considered as the size of the matrix which is directly proportional to the number of output classes.

This confusion matrix is said to be in the form of a square matrix where the column represents the True class and the row represents the predicted class of a model and it also consist of four types as:

1.3.1 False negative: Where observation is positive and model classification is negative

1.3.2 False positive: Where observation is negative and model classification is positive

1.3.3 True negative: Where observation is negative and model classification is negative

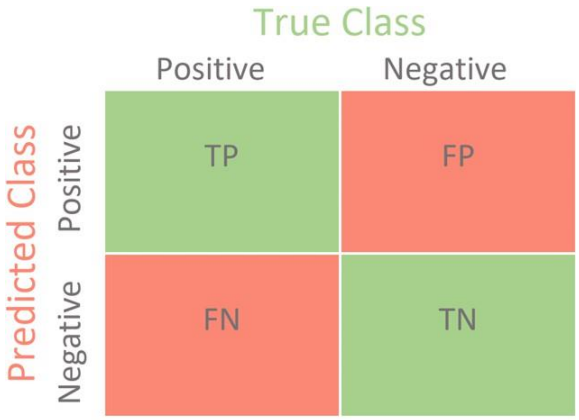


Fig 1.3: Confusion Matrix

1.4 Classification Report :

It is a performance evaluation metric in machine learning and it is used to show the precision, recall, F1-score and supports the trained classification model. It helps to provide a better understanding of the overall performance of the trained model. To understand the classification report of the machine learning model, we need to know the metrics as follows:

1.4.1 Precision:

- It is defined as the ratio of true positive to the sum of true and false positive.
- It is given as, True positives / (True positives + False positives)
- It tells how many of the correctly predicted cases actually turned out to be positive.

1.4.2 Recall:

- It is defined as the ratio of the true positives to the sum of true positives and false negative.
- It is given as, True positives/ (True positives + False Negative).
- It tells us how many of the actual positive cases are able to predict correctly with the model classification.

1.4.3 F1-score:

- It is the harmonic mean of precision and recall.

2. Literature Review

2.1 Rimpal R. Popat with Jayesh Chaudhary:

They made a survey on credit card fraud detection, considering the major areas of credit card fraud detection that are bank fraud, corporate fraud, Insurance fraud. With these they have focused on the two ways of credit card transactions i) Virtually (card, not present) ii) With Card or physically present. They had focused on the techniques which are Regression, classification, Logistic regression, Support vector machine, Neural network, Artificial Immune system, K nearest Neighbor, Naïve Bayes, Genetic Algorithm, Data mining, Decision Tree, Fuzzy logic-based system, etc. In which, they have explained six data mining approaches as theoretical background that are classification, clustering, prediction, outlier detection, Regression, and visualization. Then have explained about existing techniques based on statistical and computation which is Artificial Immune system (AIS), Bayesian Belief Network, Neural Network, Logistic Regression, Support Vector Machine, Tree, Self organizing map, Hybrid Methods, As a result, they had concluded that all the present machine learning techniques mentioned above can provide high accuracy for the detection rate and industries are looking forward to finding new methods to increase their profit and reduce the cost. Machine learning can be a good choice for it.

3. Model Building

The machine learning models need to be built because to unbalance the dataset according to its algorithm as follows:

I. Logistic Regression
II. Decision Tree classifier
III. Random Forest classifier
IV. XG Boost classifier
V. Support Vector Machines (SVM)

3.1 Logistic Regression :

It is one of the most popular machine learning algorithms which comes under supervised learning technique and also it is used for predicting the dependent variable using a given set of independent variables.

3.2 Decision tree Classifier :

It is a tree – structured classifier in which the internal nodes represents the features of the dataset, the branches represents the decision rules and the leaf node represents the outcome. It is a supervised learning technique that is used for both classification and regression problems but mostly it is mostly preferred for solving classification problems.

3.3 Random Forest Classifier:

It is a machine learning technique which is used to solve regression and classification problems and it also utilizes the ensemble learning which is a technique that combines many classifiers to provide solutions to the complex problems.

3.4 XG Boost Classifier:

It is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements the machine learning algorithms under the gradient boosting frameworks and it provides a parallel tree boosting to solve many data science problems in an accurate way.

3.5 SVM :

It is a supervised learning machine algorithm which is used for both classification or regression problems and mostly it is used as classification problems. In SVM algorithm, it plots the data item as a point in n – dimensional space with the value of each features being the value of a particular coordinate. It also performs classification by finding the hyper-plane that differentiates the two classes. Hence, these algorithms are the part of SK learn package libraries in which it includes ensemble based methods for classification and regression.

4. Cross Validation in Machine Learning

In order to balance the classes, the following methods have been taken as follows:

4.1 SMOTE :

It stands for Synthetic Minority Oversampling Technique and it is a statistical technique in order to increase the number of cases in the dataset in a balanced way. This module works by generating the synthetic data based on the feature similarities between existing minority instances. In order to create a synthetic instance, it will find the K - Nearest neighbors of each minority instances. In order to create a synthetic instance, it finds the K - Nearest Neighbor of each minority instance, randomly selects one of it and then it calculates the linear interpolation to produce a new minority instance.

5. Implementation

This dataset contains transactions made by European card holders over a period of two days in September 2013 out of which the total transaction is 2,84,807 from which 492 were fraudulent. As this data is highly unbalanced, the positive class or known to be frauds accounted for 0.172% of the total transactions. This dataset has been modified with Principal Component Analysis (PCA) in order to maintain confidentiality. Apart from time and amount, all other features from v1 to v28 are the principal component obtained using PCA.

Since it has done in google collab, it needs to get mounted with the google drive in order to generate the code by importing the csv file to contain the dataset. We have also plotted different graphs to check the transactions in the data.

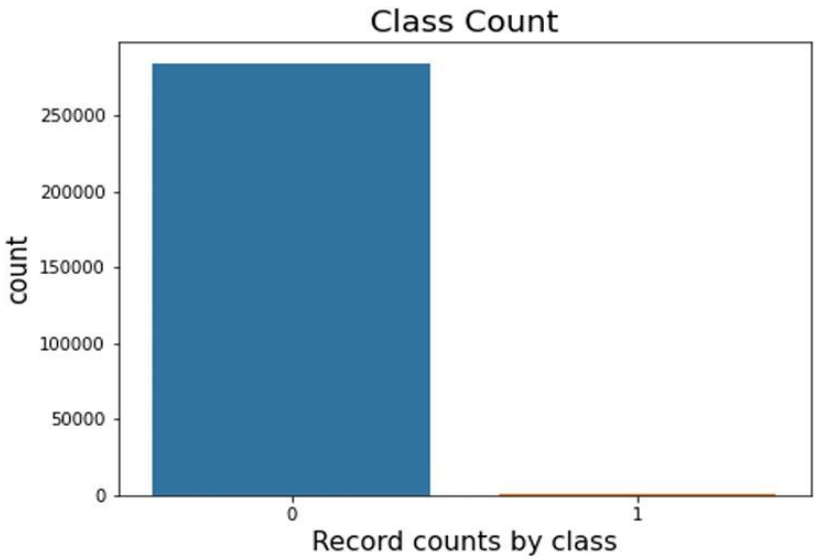


Fig 5.1: Class count

Here this graph shows the number and percentage of fraudulent and non- fraudulent transactions in a bar graph in which the fraudulent transactions are much lower than the non-fraudulent transactions.

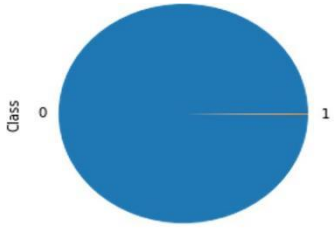


Fig 5.2: Pie Chart of Class Count

After the analysis of the data, we have to check whether the data is correlated or not, for that we need to generate a heat map in order to find weather it is strongly correlated or it is weakly correlated before splitting the data into train data and test data.

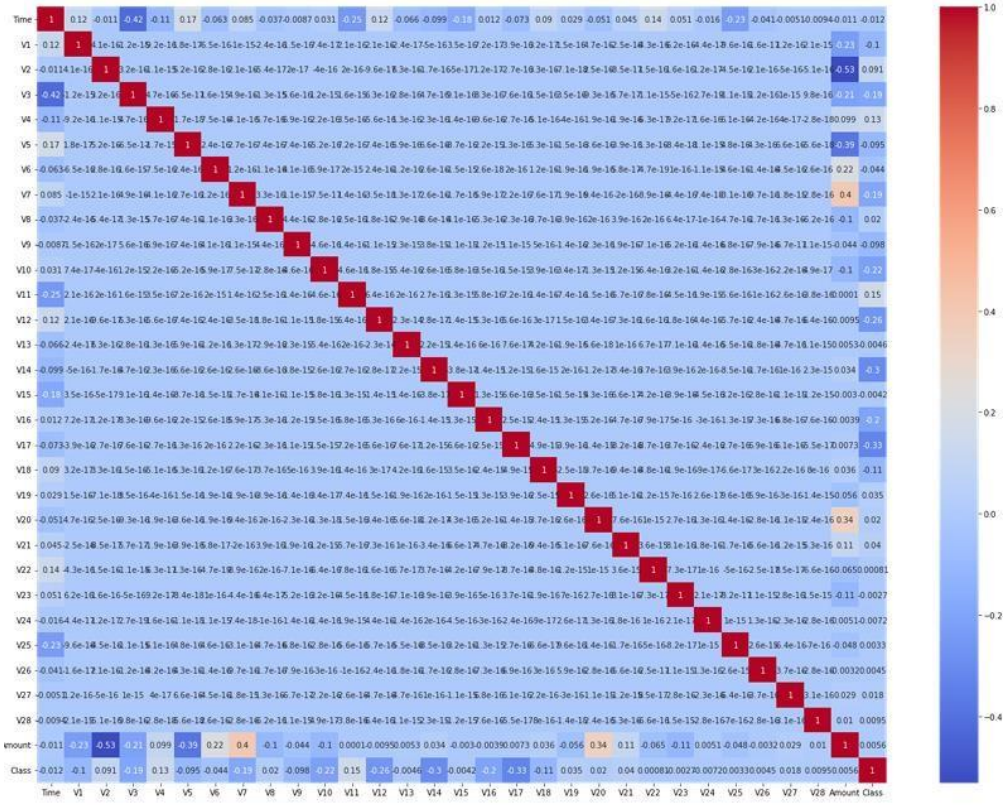
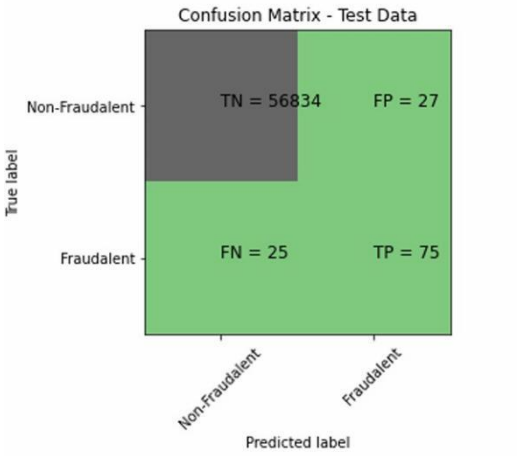


Fig 5.3: correlation in heatmap

So in this graph, the heat map is produced in order to check the correlation where the darkly colored represents that it is strongly correlated and the lightly colored represents it is weekly correlated.

5.1 Decision Tree :



classification Report:

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	56961
1	0.74	0.75	0.74	100
accuracy			1.00	56961
macro average	0.87	0.87	0.87	56961
Weighted average	1.00	1.00	1.00	56961

entropy tree roc value: 0.8747625789205256

Tree threshold: 1.0

ROC for the test dataset 87.5%

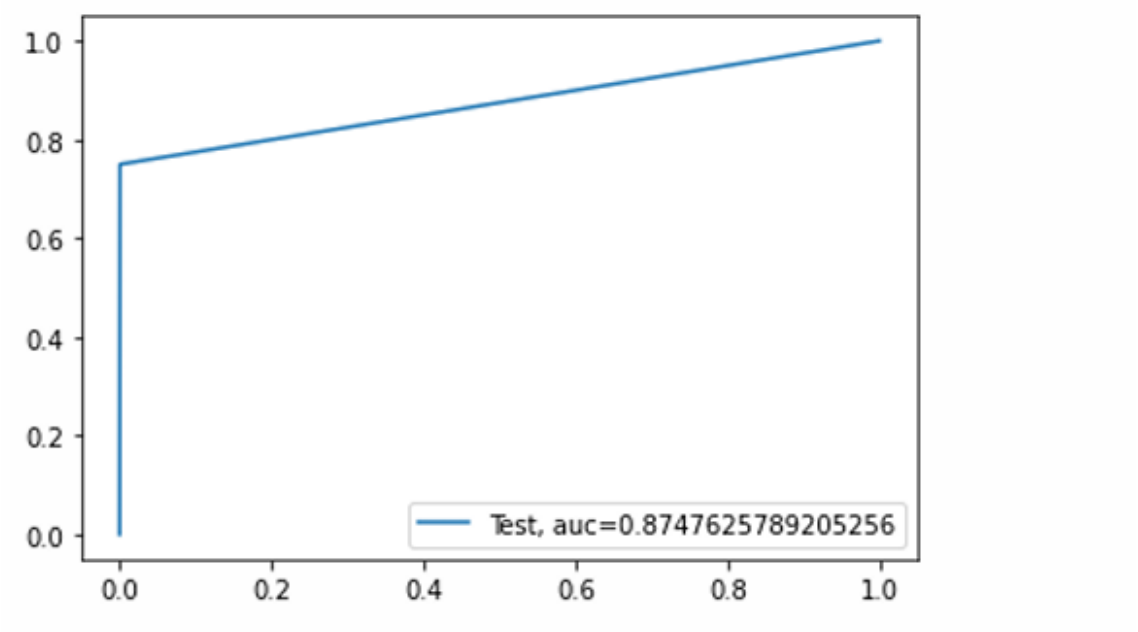
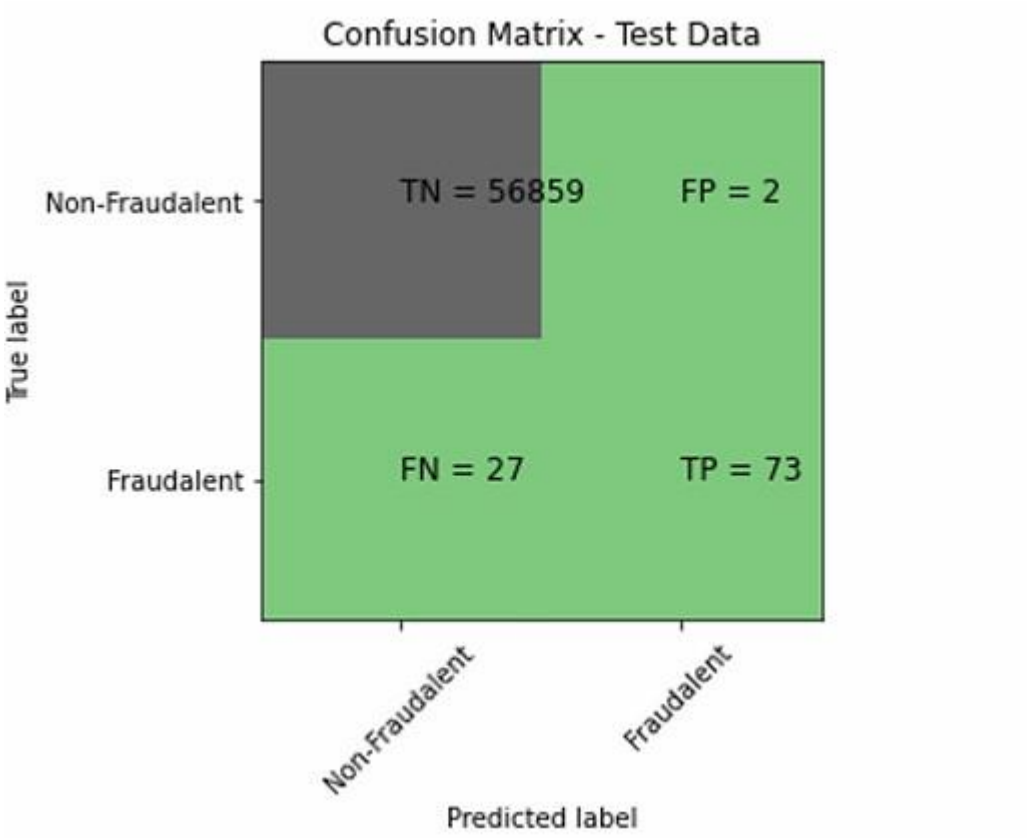


Fig 5.13: ROC curve of Decision Tree

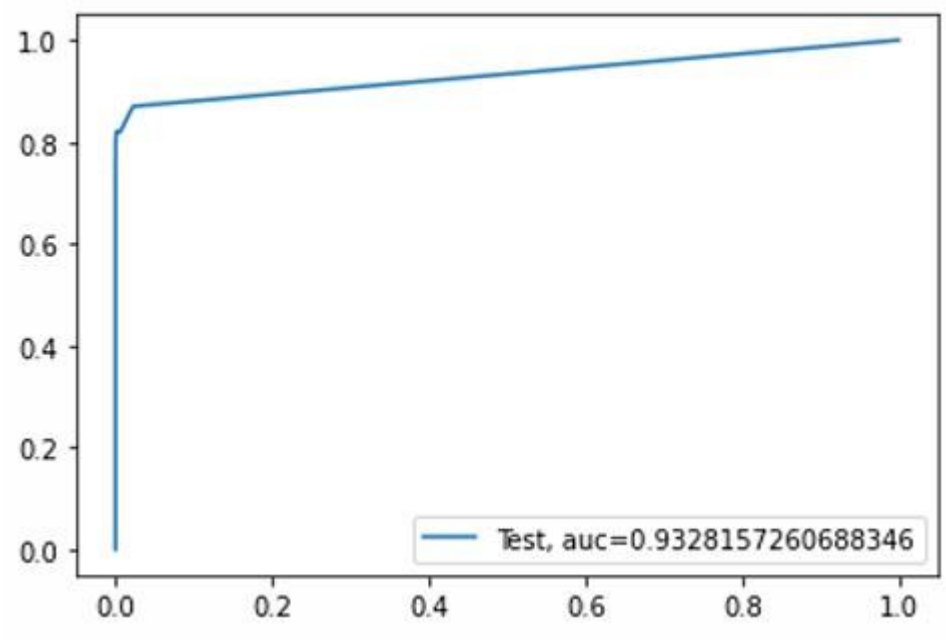
5.2 Random Forest Model :



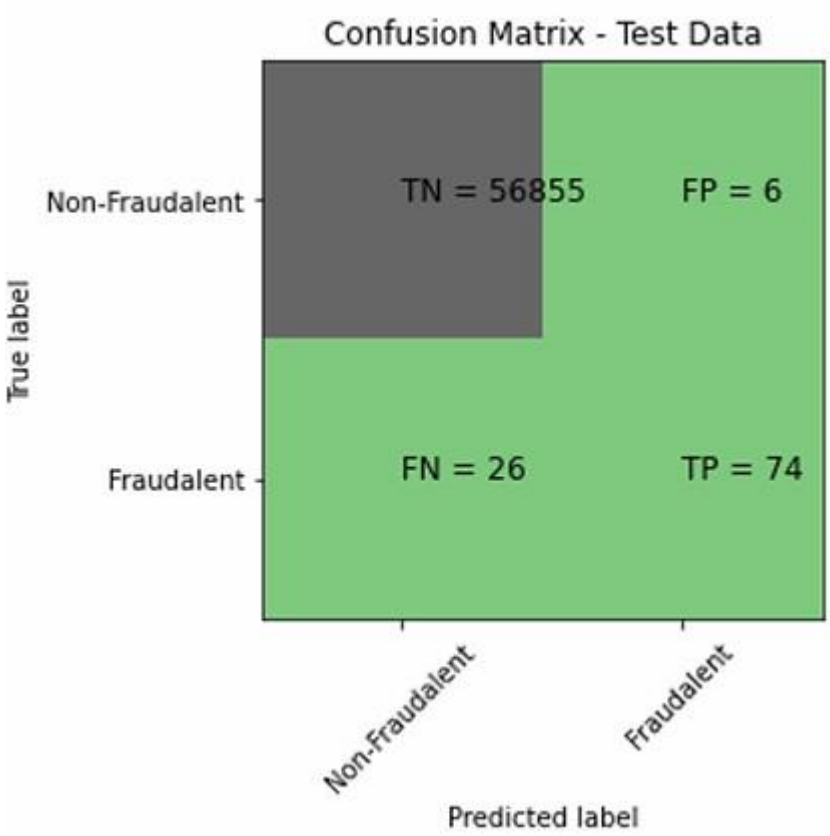
Classification Report :

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	56961
1	0.97	0.73	0.83	100
accuracy			1.00	56961
macro average	0.99	0.86	0.92	56961
Weighted average	1.00	1.00	1.00	56961

ROC for the test dataset 93.3%

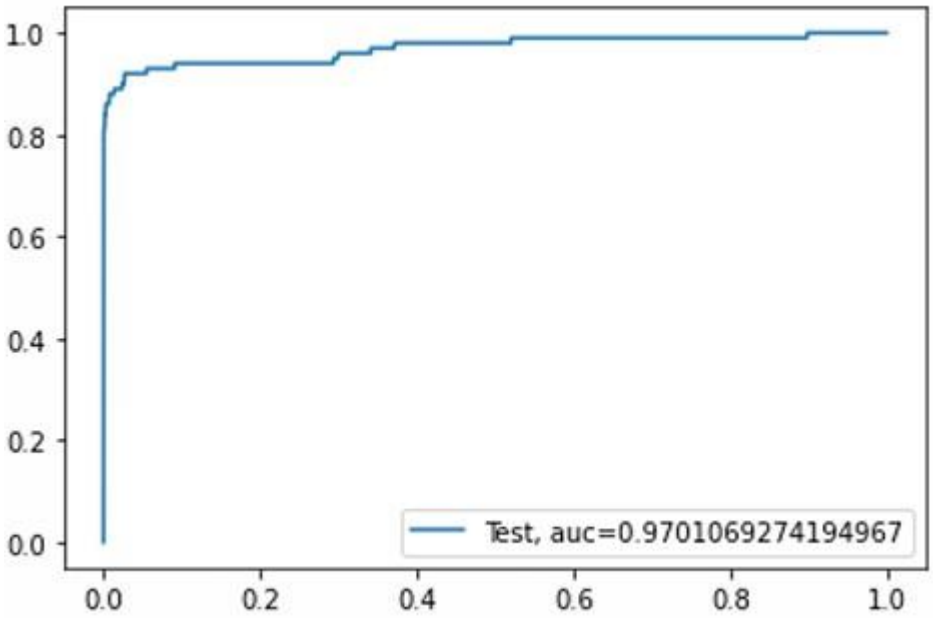


5.3 XG Boost Model :

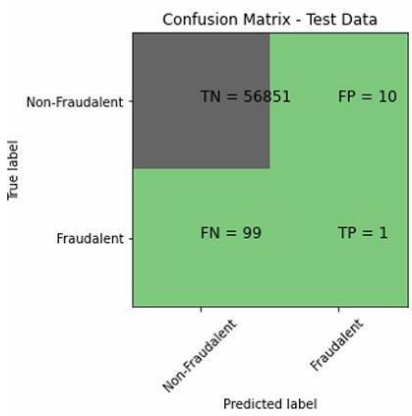


Classification Report :

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	56961
1	0.93	0.74	0.82	100
accuracy			1.00	56961
macro average	0.96	0.87	0.91	56961
Weighted average	1.00	1.00	1.00	56961

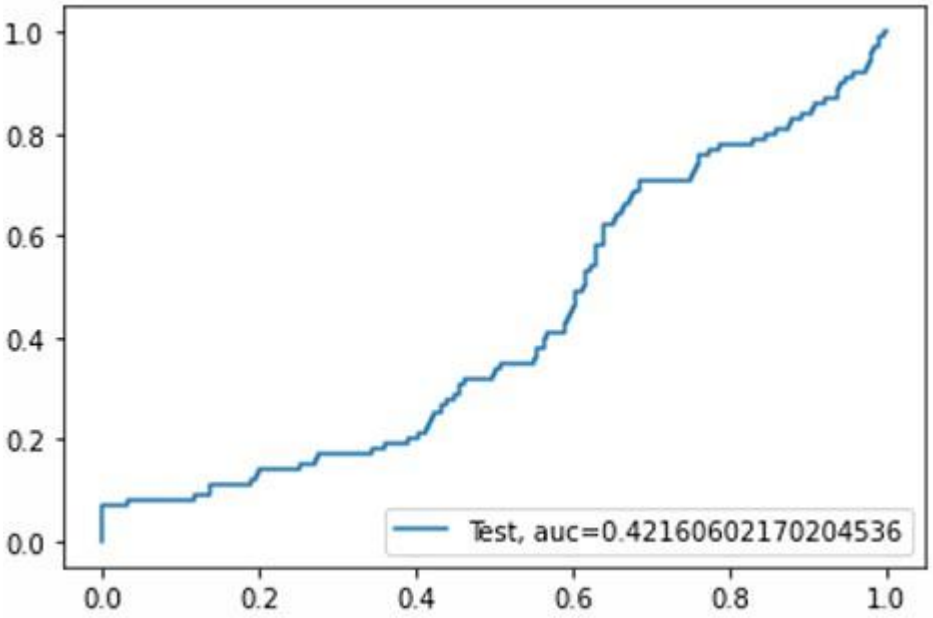


5.4 SVM Model :

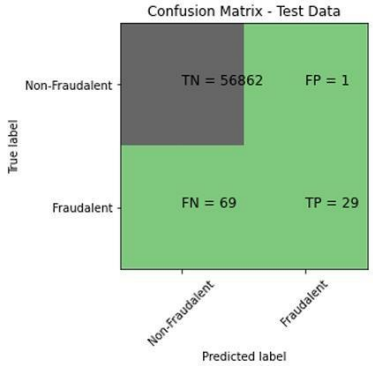


Classification Report :

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	56961
1	0.09	0.01	0.02	100
accuracy			1.00	56961
macro average	0.54	0.50	0.51	56961
Weighted average	1.00	1.00	1.00	56961

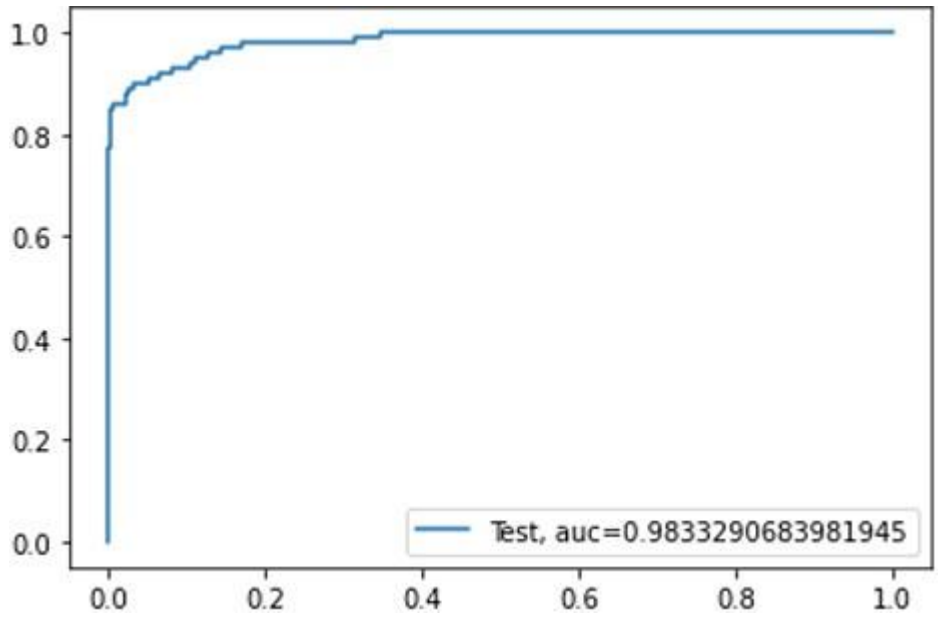


5.5 Logistic Regression :



Classification Report :

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	56961
1	0.97	0.30	0.45	98
accuracy			1.00	56961
macro average	0.98	0.65	0.73	56961
Weighted average	1.00	1.00	1.00	56961



6. Conclusion

The **Credit Card Fraud Detection** dataset is a valuable resource for building predictive models that identify fraudulent transactions. Machine learning models can be developed using this dataset, with techniques like resampling, anomaly detection, and ensemble learning used to address the class imbalance. Evaluating model performance requires specialized metrics, including precision, recall, F1-score, and ROC-AUC, due to the nature of the fraud detection problem.