Credit Card Fraud Detection

Keshav Atri (202118002)
MSc Data Science, DAIICT
Delhi, India
keshavatri20@gmail.com

Pranay Kothari (202118010)
MSc Data Science, DAIICT
Udaipur, Rajasthan, India
kotharipranay2000@gmail.com

Jainam Shah (202118014)
MSc Data Science, DAIICT
Ahmedabad, Gujarat, India
jainamshah535@gmail.com

Prachi (202118021)
MSc Data Science, DAIICT
Ahmedabad, Gujarat, India
prachikshah18@gmail.com

***Abstract*—Credit card plays a significant role in present wealth. It is important for companies to identify fraudulent credit card transactions so that they don't charge customers that they did not purchase any goods. Since the day payment systems have emerged, group of people have been trying to illegally access other people's money. As there is rapid spread of fraud in digital economy, fraud detection becomes an important issue which needs attention right away. Frauds happening due to credit cards mainly involve a payment of items or goods and services from other person's account without their consent. According to the Nilson Report, it estimates that by 2025, United States will suffer loss upto 12.5 billion dollars due to credit card fraud.**

This project uses various Machine Learning algorithms to detect Credit card fraud and find the best possible model that can detect fraud in best possible way. In this project we have used Logistic Regression, Random Forest Classifier and KNN Classifier as well. We have drawn the confusion matrix, classification report and ROC curve for all. The results of these models help us give a clear idea whether our transaction is legit or fraudulent.

I. Introduction

We see frauds occurring around us here and there in any transaction, but debit card or credit card transaction is a very common example. Any person being the victim of any such fraud is both robbed of his/her money as well as his/her mental and physical health also. Credit card fraud effects the whole chain, right from manufacturer to the customer. This type of fraud is very common but also difficult to detect. Credit card fraud varies from stealing a small amount to stealing credit cards to taking account over or even more. The amount of money transacted illegally has been increasing day by day.

II. Dataset description

The dataset used in this project is from Kaggle(creditcard.csv), which consists data of two day's transactions in European Union by the cardholders on 09/2013. Total of 284,807 transactions were made out of which 492 were fraudulent. So, our dataset is highly imbalanced, with frauds less than 0.17 percent in total.

The dataset consists only of numerical variables. Due to confidentiality reasons, all the features leaving time and amount features were converted. After applying Principal Component Analysis (PCA) on the features, we get V1, V2,..., V28. These features include, credit history, earlier months bills, credit boundary, masculinity, status of prevailing account, nuptial status, earlier months payments, wage assignments, savings account, persistence, volume of credit, possessions, employment status, age in months, housing, time, amount, etc. Class column is the target variable defining whether the transaction was fraud or not Fraudulent

III. Data Pre-processing

The original dataset needs pre-processing. The following are the steps taken to change the data so that we can get better and efficient results.

1. Standardizing:
In this project we have standardized the amount feature. The range of amount in the feature varies drastically, so we standardized it.

2. Dropping a Feature:
Time feature is an external deciding factor, so we can drop it.

3. Removing Duplicates:
There are 1,081 duplicate transactions in our dataset, so we dropped them.

4. Working on imbalanced data:
As our dataset is highly imbalanced, we have applied both under sampling as well as over sampling on our dataset.

A. Under sampling:
We have 473 fraudulent transactions in total. We have randomly selected 473 legit transactions out of 284315 transactions and applying models on the new dataset.

B. Over sampling:
We have only 473 fraudulent transactions in total. In over sampling technique, it creates copies of the data. We used SMOTE technique and created 284315 fraudulent transactions and then applied models on the new dataset.

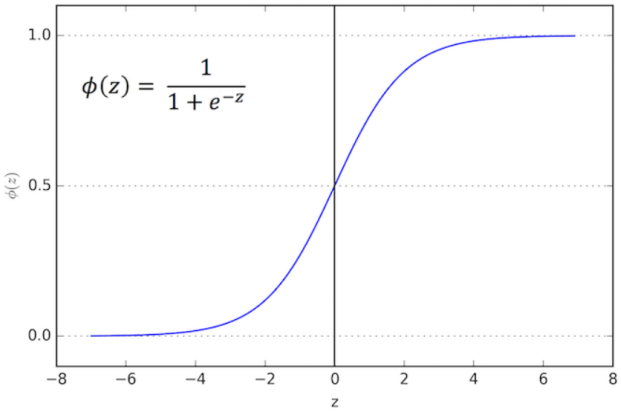
5. Splitting of dataset:
We have divided the whole dataset into training and testing sets. The training set consists of 80 percent of the dataset and testing set consists of remaining 20 percent

IV. Classifications

We used different libraries and tools along with different classifiers models on the dataset. We even compared each classifier based on its accuracy, precision, recall and f1 score.

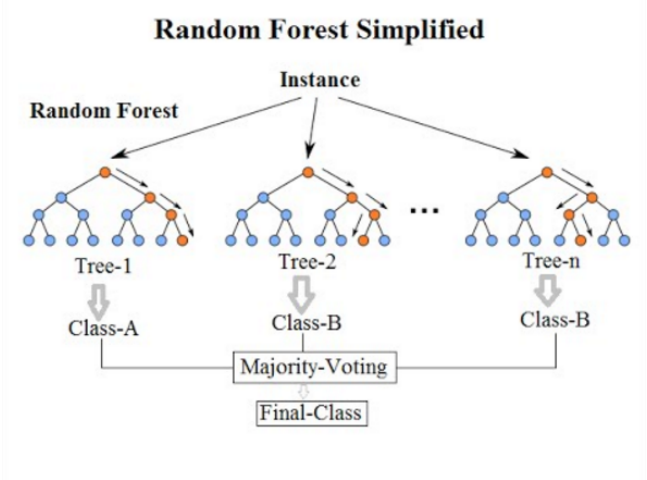
A. Logistic Regression:

Logistic Regression is one of the most commonly used classification algorithms. Logistic Regression is used when the dependent variable is categorial. Sigmoid function is applied on the predicted values to get the probability. As probability ranges from 0 to 1, our result from this regression will range between 0 to 1. According to the result, we classify to either of the class.



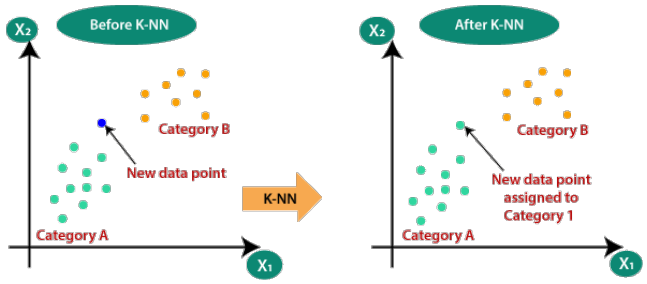
B. Random Forest Classifier:

Random Forest Classifier is a kind of ensemble classifier. It uses decision tree algorithm in randomized way. The whole dataset is converted to boot strap dataset. Based on the total number of features, we randomly make multiple decision trees and output is generated. If majority of the output is 0 then 0 is the output or vice versa.



C. KNN Classifier:

KNN represents K nearest neighbor Classification. Based on Euclidean Distance, K number of neighbor predicted values are taken to consideration. In our project we have taken 5 nearest neighbors into consideration. Based on the majority of the output of those 5 values, we declare the result.



V. Evaluation of Classification Models

For evaluating our classification model, we have used Accuracy, Precision, recall and F1 score as our parameters.

A. Accuracy:

Accuracy is to find whether the given model is best fitted or not for finding true predictions, patterns and relationships between variables.

$$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{All Samples}}$$

B. Precision:

Precision is the quality of positive predictions made by the model. Exact positive refers to the number of true positives divided by the total number of predictions.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

C. Recall:

Recall tells us about the percentage of classes that we are focused on which were reproduced by the model.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

D. F1 Score:

F1 score is represented by the weighted average of precision and recall, where 1 indicates the best and 0 indicates worst

score.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

E.ROC Curve:

A ROC Curve represents Receiver Operating Charactristic Curve. It is a graph showing the performance of the classification models. This type of plot has two parameters, True Positive Rate, False Positive Rate.

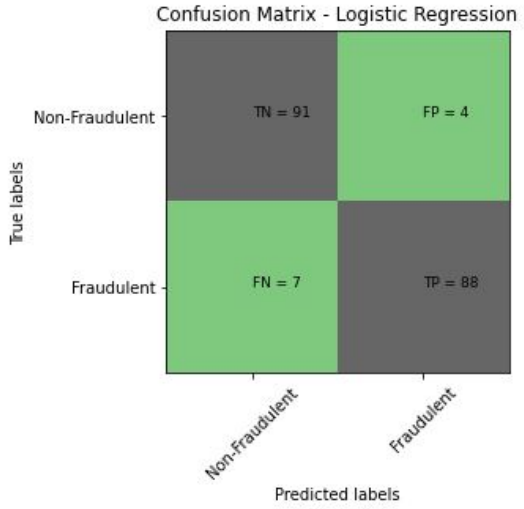
VI. Experimental Results

In our Credit Card Fraud Detection project, we have used three models, Logistic Regression, Random Forest Classifier and KNN Classifier. We have fitted our model twice, one for under sampling and over sampling. We further compared our models and found out which is the most suitable based on the accuracy, precision, recall, F1 score.
The experimental results are as follows :

A. Under sampling:

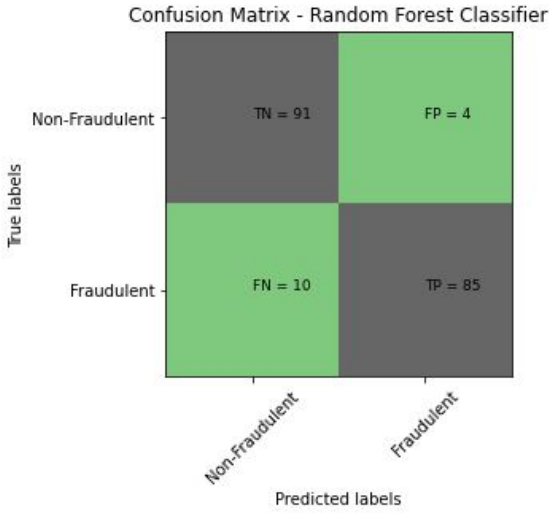
Logistic Regression

0	Logistic Regression	0.942105	0.956522	0.926316	0.941176



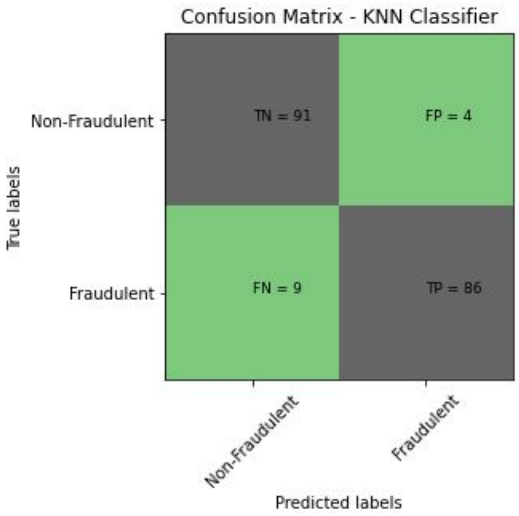
Random Forest Classifier

1	Random Forest Classifier	0.926316	0.955056	0.894737	0.923913



KNN Classifier

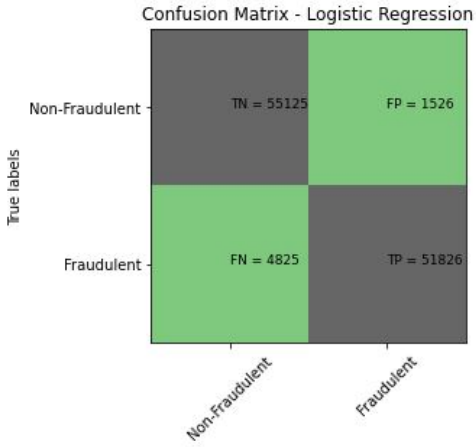
2	KNN Classifier	0.931579	0.955556	0.905263	0.929730



B. Over sampling:

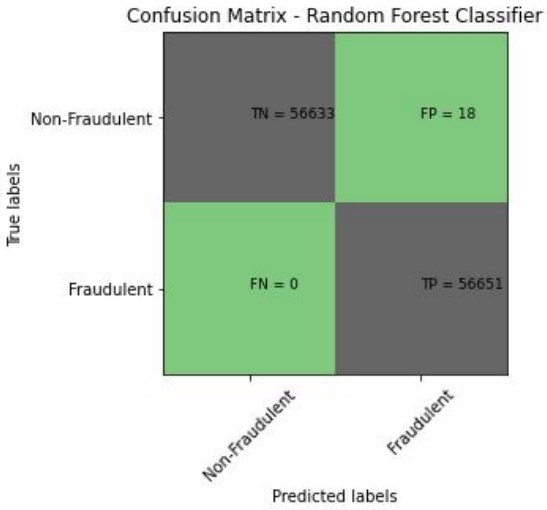
Logistic Regression

0	Logistic Regression	0.943946	0.971398	0.914829	0.942265



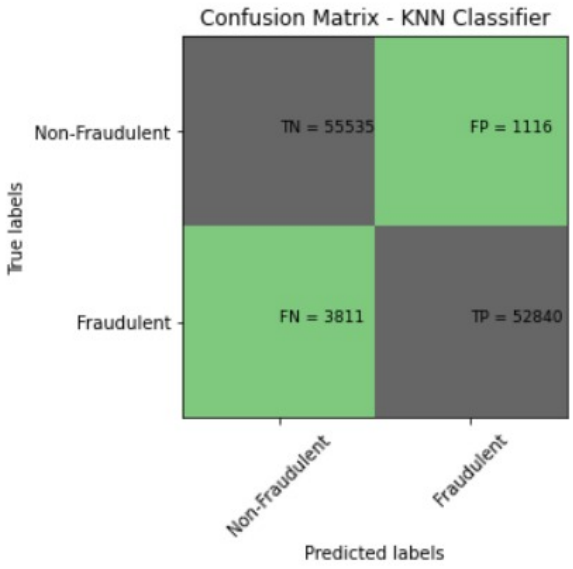
Random Forest Classifier

1	Random Forest Classifier	0.999841	0.999682	1.000000	0.999841



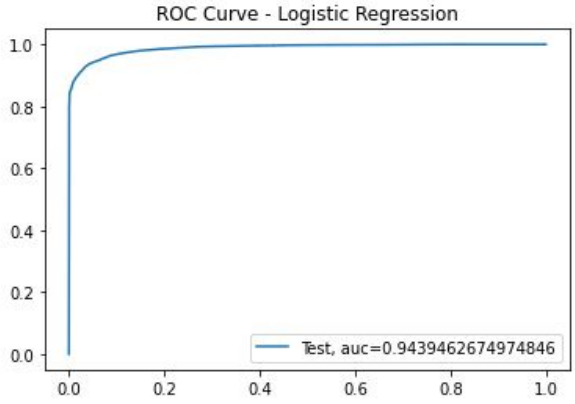
KNN Classifier

2	KNN Classifier	0.956514	0.979316	0.932728	0.955455

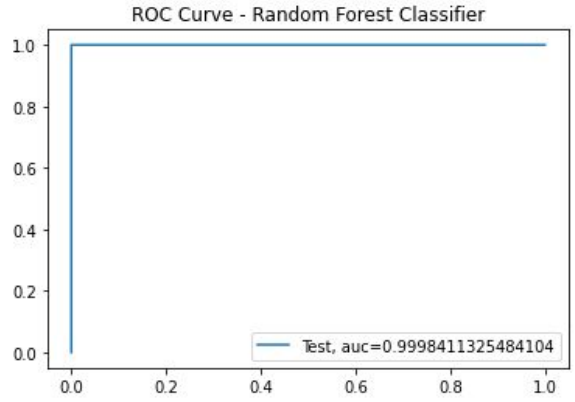


C. ROC Curve :

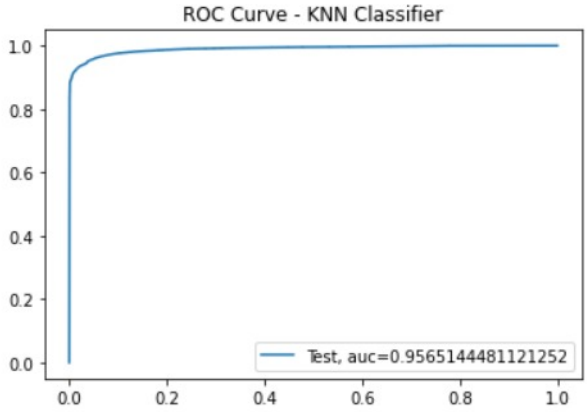
Logistic Regression (Highest Accuracy ROC Curve)



Random Forest Classifier (Highest Accuracy ROC Curve)



KNN Classifier (Highest Accuracy ROC Curve)



VII. Conclusion

In this project, we have two types of data preprocessing, namely Under sampling and Over sampling. We have applied three models for both Under sampling and Over sampling. In the case of Under sampling, Logistic Regression has the highest accuracy of 0.942105 with precision 0.956522. In case of Over sampling, there is an improvement in the accuracy of Random Forest Classifier with the highest accuracy of 0.999841. In the case of precision also, we see that Over sampling is better than under sampling.

In Under sampling, we cut down the values to the minimum requirement which effects our results. In the case of over sampling we have large data to process with and it gives us better results.

VIII. Future Work

From the above analysis, we saw that different algorithms of machine learning are used to detect if a transaction is legit or fraud but the results are not up to the mark. We can also try implementing genetic algorithms and different types of stacked algorithms. We would like to implement deep learning algorithms that could detect fraudulent transactions more accurately.

IX. References

1.https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/

2.https://towardsdatascience.com/dealing-with-imbalanced-dataset-642a5f6ee297

3.D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, ”Credit Card Fraud Detection Using Machine Learning,” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 967-972, doi: 10.1109/ICICCS51141.2021.9432308.

4.F. Ahmed and R. Shamsuddin, ”A Comparative Study of Credit Card Fraud Detection Using the Combination of Machine Learning Techniques with Data Imbalance Solution,” 2021 2nd International Conference on Computing and Data Science (CDS), 2021, pp. 112-118, doi: 10.1109/CDS52072.2021.00026.

5.R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, ”Credit Card Fraud Detection Using Machine Learning,” 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114.

6.J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, ”Credit card fraud detection using machine learning techniques: A comparative analysis,” 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.