International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# A Study on Detection of Credit Card Fraud using Machine Learning Techniques

**Pranjul Sharma**

Software Engineer, Radisys

**Abstract**

Credit Card Fraud, a fraudulent activity committed by stealing a credit card and using the same without knowledge or permission of the Card owner. Credit Card Fraud is usually committed by a criminal to purchase goods or services utilizing another account but the same card. Machine Learning algorithms can simplify the process of detection of fraudulent transaction. We show how various algorithms can be used to determine if the transaction is legitimate or not. We have split the dataset into test and train data. SMOTE technique has been on the train data for oversampling as the dataset being used is highly imbalanced. Machine Learning Classification techniques such as Gaussian Naive Bayes, Logistic Regression and Random Forest have been compared in this research. Based on the obtained performance measures, we can conclude that all three Machine Learning Models could be used for Fraudulent Transaction Detection.

**Keywords**: Logistic Regression; Random Forest; Gaussian Naive Bayes; SMOTE; Credit Card Fraud

## I. INTRODUCTION

In today's world, vast amounts of data are being processed daily, and all companies around the world are trying to provide their customers with best services. This data is their business prospect in the coming future, it needs to be stored for processing. More importantly it needs to be stored in the most secure way possible. These financial institutions collecting data must ensure that their customer data is safely and securely available only to their organization. Security protocol must be followed as Data breach could have a serious impact on the company as well as its customers' financial assets. These institutions must also remember that their public reputation is a big factor in their well doing, data breach would cause damaged brand reputation [1]. In most of the cases involving security breach, financial information is stolen. When an individual has to pay for goods and services that they may not have willingly opted for, it classifies under Financial Fraud. Frauds such as identity theft, investment fraud, credit card fraud, and insurance fraud also come under Financial Frauds. In today's digital era, cashless transactions have made payments much more straightforward and have rapidly grown popular. Due to this, credit card fraud has become widespread. When a criminal, uses another person's credit card to purchase goods or services for personal needs without knowledge of the card owner, it is referred to as Credit Card Fraud. Sometimes, these frauds can be a very organized crime.

Many fraud customer care centers, fool customers into believing they are paying for the customer care service or a product, these criminals fool customers into giving their private credit card numbers and they sometimes

even fool the customers to provide the secure One Time Passwords, claiming they have won some car or other gifts.

Credit Card frauds in the year 2018, in the United Kingdom alone came up to a total close to 850 million pounds. Credit Card companies have been able to prevent 1.6 billion pounds of fraudulent transactions in the year 2018. Companies have developed innovative and sophisticated methods to avoid these, yet fraudsters create new ways to gain access.

Machine Learning methods are thus implemented to categorize transactions as fraudulent or not to prevent fraudsters from spending large amounts of money before the cardholder is aware of it.

This Research Paper studies the performance of different Machine Learning techniques, namely Gaussian Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR), and comes to a conclusion on which techniques is most appropriate for the Fraudulent Credit Card Transaction Detection.

This Paper is ordered in the following manner: Related Works, discussion of the various research performed on this topic, followed by Methodology, a brief discussion about the Machine Learning methods to be applied on the dataset and its description. Experimental Results and Analysis, involves the study of the obtained Results. Followed by Conclusion, Acknowledgement and References

## II. RELATED WORKS

Major losses were being incurred due to these fraudulent activities which inspired researchers across the world to find a solution eradicate this problem. A few methods have been developed and tested and are discussed below.

Techniques such as Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), Logistic Regression (LR) as well as a few other Classification Techniques were implemented in paper [2] resulting in High Recall. Upon subjecting the data to under-sampling, F1 score saw a great increase, due to the increase in Precision.

In paper [3], the same dataset was used and comparison was made between the models based on RF, LR and DT. Of the three techniques, RF came out on top of the results, obtaining an accuracy score of 95.5%, LR and DT followed with respective accuracy scores of 90% and 94.3%.

Papers [4] and [5], show that both techniques k-Nearest neighbors (KNN) and Outlier Detection are quite helpful in detection of Fraudulent Credit Card Transactions. KNN was also used in paper [6] and was found to perform well. In paper [7] a comparison was made between deep learning techniques and classical algorithms. An accuracy score of 80% was achieved using Deep Learning Technique.

In paper [8], algorithms such as :- Gradient Boost, Logistic Regression, Support Vector Machine, DT, K Nearest Neighbors , Random Forest, XGBoost (XGB), Multi-Layer Perceptron and stacking classifier were implemented. All of the above-mentioned algorithms achieved an accuracy over 90% with stacking classifier being the highest with an accuracy and recall of 95%.

Authors of Paper [9] and [10] demonstrated that the use of Neural Networks can show an improvement in results.

Paper [11] shows how MLP could be used for the Detection of Fraudulent Credit Card Transactions. This conclusion was obtained after a comparative study of Restricted Boltzmann Machine Algorithm, MLP and Auto-Encoder which were tested on three different datasets.

In most of these papers authors have implemented under- sampling technique and this paper uses a different approach of oversampling called SMOTE, with a primary aim of showing how SMOTE could help boost the predictive nature of different machine learning algorithms can give optimal results. In this research we are also evaluating the performance of the techniques using Area Under the Curve of the Receiver Operating Characteristic Curve which is a plot with the True Positive Rate on y-axis and False Positive Rate on x-axis. The Results obtained would be compared after subjecting the dataset to the following Machine Learning Techniques, NB, LR and RF, which would be used for the Detection of Fraudulent Credit Card Transactions.

## III. METHODOLOGY

### A. Dataset

The dataset was obtained from Kaggle, Credit Card Fraud Detection. Credit Card Transactions made by European Credit Card Holders in the year 2013. The transactions were collected over 2 days [12].

There are exactly 284,807 credit card transactions in the dataset, of which only 492 transactions were fraudulent. Only 0.173% of transactions are labelled as fraudulent, we can say the dataset is highly unbalanced. Due to confidentiality reasons, we are given 28 numeric features without any details about their origin. The original input features had undergone PCA transformation.

Apart from these 28 numeric features, we are provided with two more independent features namely, “Amount” and “Time”. The dependent variable is called “Class”, wherein value 1 signifies a fraudulent class and value 0 signifies a normal transaction. Amount refers to the transaction Cost.

Variable Time refers to the time (in seconds) of the Transaction.

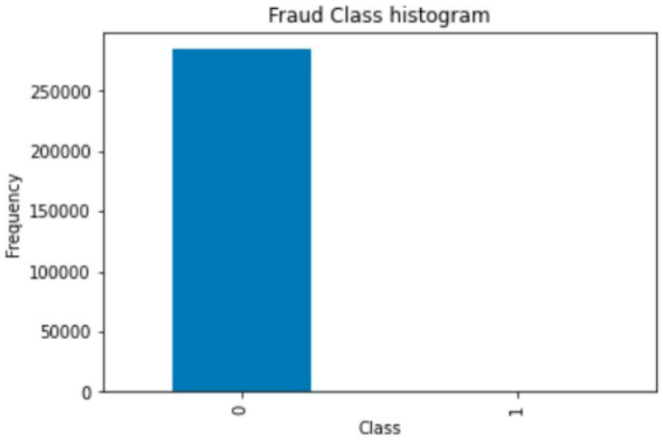
### B. Preprocessing

Initially, the step of Data Preprocessing involves Data Cleaning, wherein we fill in the missing values for different features. Missing values would not only increase the time and complexity while training the model, it would also cause bias in the model. Dataset used in this Research did not contain any missing values. Features that did not have any meaning to the prediction such as the ‘Time” of the Transaction was removed. Time would not have any impact in our study, as fraudulent transactions could occur at any time of the day, their significance towards the classification is quite low.

Normalization, another key step during preprocessing, normalization is very important as it would scale down all the independent feature values between -1 to 1. Thereby reducing the partiality to a feature with higher numeric values in other words the model would be unbiased to the extent of numeric values. Since we have obtained V1 to V28 features through PCA they have already been normalized. We have applied the Normalization to the Feature “Amount”.

Machine Learning models tend to struggle when faced with a highly unbalanced dataset such as the one in our Research. In a Dataset if over 99% of the values are of Non-Fraud Class, even if the model predicts every data point as Non-Fraud, accuracy would be over 99%. At a glance, accuracy may be quite high which would make us think it is a very well-trained model, but in reality, the Model would not be able to predict when a Fraudulent Activity occurs which defeats the purpose of this research. Hence, we have decided to apply the Synthetic Minority Oversampling Technique which is one of the most popular oversampling methods applied to a highly unbalanced dataset.

**Figure 1. Frequency of Class prior to sampling**




**C. Experiment**

Logistic Regression, was the first model, it is the most widely used Machine Learning Classification Technique. Logistic Regression is capable of modeling the relationship between continuous, categorical, or binary independent variables to the dependent variable. Dependent Variables are usually categorical. Logistic Regression classifies depending on the probability of the given data being either one of the classes. Logistic Regression computes a threshold value upon training the model and analyzing the data, if the probability is greater than the threshold then it is classified as +ve class and below that threshold, it is classified as -ve.

Naïve Bayes, a classification machine learning technique, it is developed based on popular Bayes Theorem in Probability. Naïve Bayes is a one of kind Algorithm as it does not have any dependencies between the different features. Naïve Bayes has many different algorithms based on the type of distribution such as Gaussian, Multinominal, and Bernoulli. We would be using Gaussian Naïve Bayes for this experiment to predict fraudulent transactions. Gaussian Naïve Bayes is primarily used when features are continuous. Features fed into a Gaussian Naïve Bayes model are assumed to follow a Normal distribution.

Random Forest, a machine learning technique that is used for Classification and Regression. Random Forest involves the building of multiple Decision Trees. Random Forest generally yields better results with more the number of trees in the forest. Prevention of Overfitting of the model is something that must be maintained with Random Forest. The individual Decision Trees provide results, which may be further merged to yield better predictive values. In this experiment, we have implemented Random Forest Classifier, it would be used to classify a transaction as fraudulent or not.

We have used GridSearchCV to identify the best parameters for Logistic Regression and Random Forest. Gaussian Naïve Bayes does not have any hyperparameters. Hence, we did not need to use GridSearchCV for Naïve Bayes.

Upon using GridSearchCV and finding the best fitting parameters to maximize AUC, we were able to model our Logistic Regressor and Random Forest. The model was trained using the data after SMOTE. This would help the model in the identification of fraudulent transactions.

It was important for us to use the dataset after SMOTE, as the given dataset did not contain more than 0.18% fraudulent transactions. This could lead to what is called bias to a particular class. Such a model would have never seen the important characteristics of fraudulent activity, would therefore miss out on classification of

the same when an actual Fraud Transaction occurs. Study involving Sampling was key to our Research as the study of the working of models upon over sampling is not quite extensive. Our primary objective in this experiment was to train the Machine Learning Models, on the over-sampled data, was.
Train and Test has been split in the ratio of 70 is to 30. The training data was later sampled to obtain another sampled training data. This Sampled Data was used to train all the three models.
The Obtained Models after training were fit on the test data to obtained predicted values.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order determine which one of the algorithms is able to detect fraudulent transactions, algorithms are compared using
different performance measures. The Performance Measures which are commonly used such as Accuracy, Recall and Precision have been determined. These measures can be found easily using the Confusion Matrix.

Fig. 1.
Example of a Confusion Matrix and Formula for Precision, Recall and F1 Score

Fig. 1. Example of a Confusion Matrix

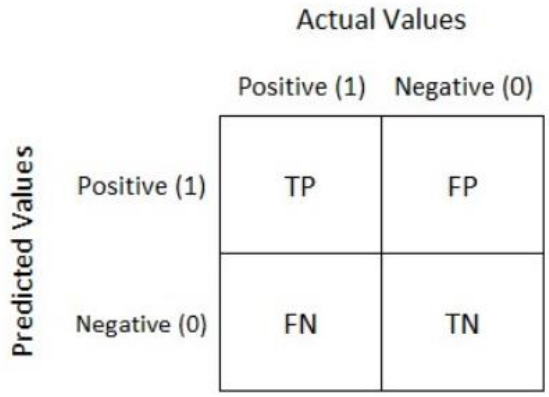



Fig. 2. Formula for Precision, Recall and F1 Score

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

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

All the models have performed much better after subjected to the sampled data when compared to the original test data. This shows us the importance of sampling for unbalanced datasets.

Results obtained on Logistic Regression Model: (Table 1)

TABLE 1: Logistic Regression Confusion Matrix

| | | *Predicted* | |
|---|---|---|---|
| | | 0 | 1 |
| *Actual* | 0 | 83581 | 1727 |
| | 1 | 16 | 119 |

- Recall: 88.14%,
- Precision: 6.44%,
- F1 Score: 0.12,
- Accuracy:97.96%.

Results Obtained on Random Forest Model: (Table 2)

- Recall: 80.00%,
- Precision: 85.03%,
- F1 Score: 0.81,
- Accuracy: 99.94%

TABLE 2: Random Forest Confusion Matrix

| | | *Predicted* | |
|---|---|---|---|
| | | *Normal* | *Fraud* |
| *Actual* | *Normal* | 85291 | 17 |
| | *Fraud* | 26 | 109 |

Results obtained on the Gaussian Naïve Bayes Model: (Table 3):

- Recall: 82.96%,
- Precision: 5.67%,
- F1 Score: 0.106,
- Accuracy: 97.79%.

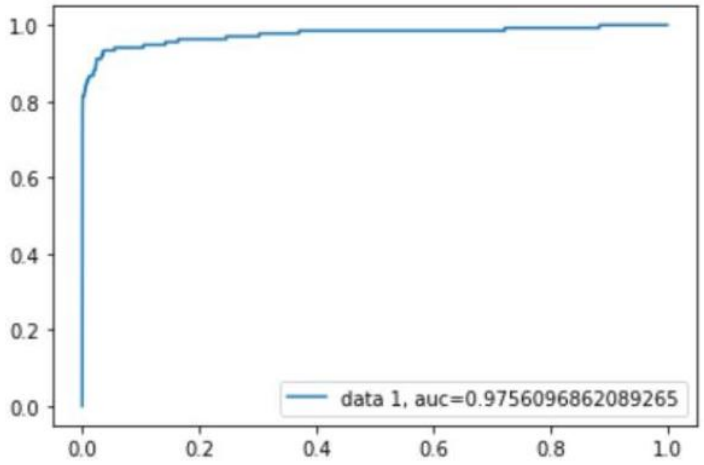
TABLE 3: Gaussian Naïve Bayes Confusion Matrix

| | | *Predicted* | |
|---|---|---|---|
| *Actual* | | 0 | *Fraud* |
| | *Normal* | 83447 | 1861 |
| | *Fraud* | 23 | 112 |

From the results, it is clear that all the three models have extremely high accuracy. High Accuracy does not imply that our results are perfect. Accuracy is merely one of the many performance measures which would be used to interpret the performance of the model. F1-Score and Area under the Curve value of the ROC curve. AUC of the ROC curve, is a performance metric which would find out the probabilities of the different classes of a classification problem and plot a graph which is a plot with the True Positive Rate on y-axis and False Positive Rate on x-axis. The results obtained in this research when compared to the other researches using same dataset, we can clearly notice a better predicting model. Oversampling has proven helpful in this research as it yields better results. AUC values of the models were quite low when subjected to the same models when trained using the unsampled dataset. This implies that Sampling is quite an important step to be dealt with in case of unbalanced datasets. Papers [3], [6] help us in understanding the differences in the results when subjecting the dataset to oversampling. Papers [9] and [13] show us how deep learning algorithms are implemented to detect fraudulent transactions and how they are helpful as the dataset is quite large. But the ease of implementation of regular machine learning models is one of the advantages in our research. Papers [15] has a very similar case study involving the SMOTE technique but the comparison of the models was restricted to Accuracy, Precision, Recall also known as the preliminary performance metrics. Use of Area Under the Curve for the comparison is quite important as it helps us catch those models which sometimes get lucky in that particular test set. Regular Machine Learning Algorithms are much easier to implement, interpret and are also financially cheaper [14].

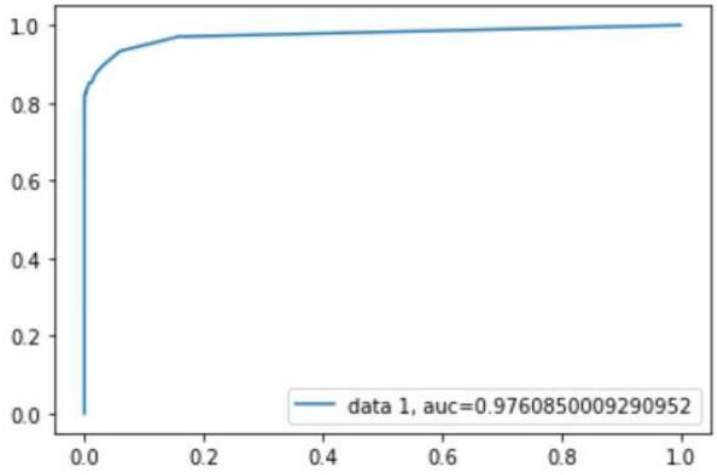
**Below are the ROC Curves of the Models:**

Logistic Regression:



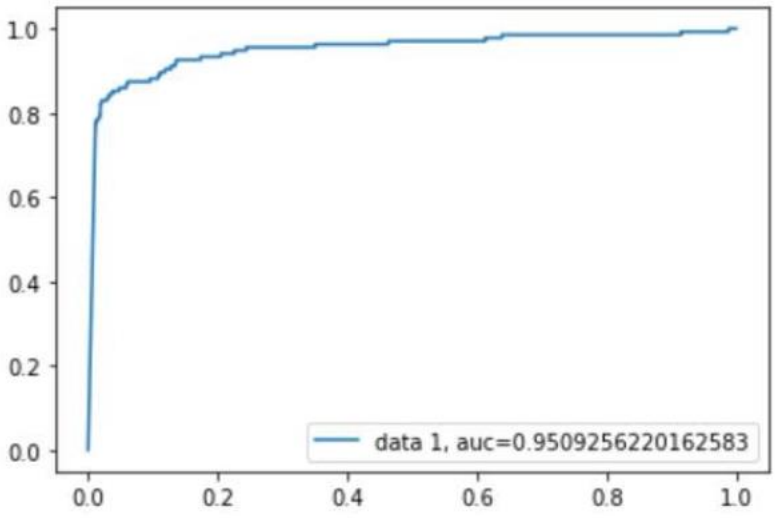

AUC Value: 0.975

Random Forest:




AUC Value: 0.976

**Gaussian Naïve Bayes:**




AUC Value: 0.95

## V. CONCLUSION

This research involved application of three different machine learning classification algorithms to detect Fraudulent Credit Card Transactions. It is a very a serious issue and negligence can lead to huge amount of monetary losses. Companies have made it an important goal to develop innovative and effective ways to tackle this issue.

Hence, upon comparing the different techniques, Random Forest proved to be the best in detecting whether transactions were genuine or not. Different metrics like F1 score and AUC (Area under curve) were used to determine the results. Random Forest obtained highest AUC, F1-Score hence it would provide the best results.

## ACKNOWLEDGMENT

We take this opportunity to thank Dr. Ruchika Malhotra for her valuable guidance throughout the research project. We value the opportunity provided to us by Delhi Technological University, DTU as well as give us the necessary resources to complete this research project.

## REFERENCES

1. Importance of Data Security for Businesses https://www.infomaxoffice.com/importance-of-data-security-for- businesses/
2. A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.
3. S. V. S. S. Lakshmi, S. D. Kavilla “Machine Learning For Credit Card Fraud Detection System”, unpublished

4. N. Malini, Dr. M. Pushpa, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection", Advances in Electrical, Electronics, Information, Communication and BioInformatics (AEEICB), 2017 Third International Conference on pp. 255258. IEEE.
5. Mrs. C. Navamani, M. Phil, S. Krishnan, "Credit Card Nearest Neighbor Based Outlier Detection Techniques"
6. J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", Computing Networking and Informatics (ICCNI), 2017 International Conference on pp. 1-9. IEEE.
7. Z. Kazemi, H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions", Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4th International Conference on pp. 630-633. IEEE
8. S. Dhankhad, B. Far, E. A. Mohammed, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study", 2018 IEEE International Conference on Information Reuse and Integration (IRI) pp. 122-125. IEEE
9. N. Kalaiselvi, S. Rajalakshmi, J. Padmavathi, "Credit card fraud detection using learning to rank approach", 2018 Internat2018 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC) ional conference on computation of power, energy, Information and Communication (ICCPEIC) pp. 191-196. IEEE
10. F. Ghobadi, M. Rohani, "Cost Sensitive Modeling of Credit Card Fraud using Neural Network strategy", 2016 Signal Processing and Intelligent Systems (ICSPIS), International Conference of pp. 1-5. IEEE.
11. A. Pumsirirat, L. Yan, "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine", 2018 International journal of advanced computer science and applications, 9(1), pp. 18-25
12. Kaggle.com. (2019). Credit Card Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud [Accessed 10 Jan. 2019].
13. Learning – Towards Data Science. [online] Available at: https://towardsdatascience.com/deep-learning-vs-classical- machinelearning-9a42c6d48aa [Accessed 19 Jan. 2019].
14. Deeplearningbook.org. (2019). Deep Learning. [online] Available at: https://www.deeplearningbook.org/ [Accessed 11 Jan. 2019].
15. Dejan Vermedja, Mirjana Karanovic, Srdian Sladojevic, Marko Arsenovic, Andras Anderla, "Credit Card Fraud Detection – Machine Learning Methods", 18th Internation Symposium Infoteh-Jahorina, 20- 22 March 2019