iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33547

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Fraud Detection in Credit Cards with Machine Learning

Anjali Chouksey[1], Riya Nimje[2], Jahanvi Saraf[3]
[1, 2, 3]MPSTME, Shirpur

Abstract: Online transactions have increased dramatically in this new 'Social-distancing' era. With online transactions, Fraud in online payments has also increased significantly. Frauds are a significant problem in various industries like insurance companies, baking, etc. These frauds include leaking of sensitive information related to the credit card which can be easily misused. Due to the government also pushing online transactions, E-commerce is on a boom. But due to increasing frauds in online payments, these E-commerce industries are suffering a great loss of trust from their customers. These companies are finding Credit card fraud to be a big problem. People have started using online payment options and thus are becoming easy targets of credit card fraud. In this research paper, we will be discussing machine learning algorithms. We have used Decision tree, XGBOOST, k-nearest neighbour, Logistic-regression, Random forest and SVM on a dataset in which there are transactions done online mode using Credit cards. We will test all these algorithms for detecting fraud cases using the confusion matrix, F1 score, and calculating the accuracy score for each model to identify which algorithm can be used in detecting frauds.

Keywords: Machine Learning, Fraud Detection, Artificial Intelligence, Decision Tree, K Nearest Neighbour, Random Forest, XGBOOST, Logistic Regression, Support Vector Machine

I. INTRODUCTION

The pandemic that was caused due to the outbreak of corona-virus is not yet over. This pandemic has changed a lot of things. Due to the out-break, Lockdown was imposed in many countries. Since lockdown, people started to switch from traditional purchasing methods to online purchasing. Now with increasing online payments mode made available for almost anything and everything, the risk of online fraud has also increased. Hacking and Phishing have increased significantly. Attempts are being made to get customer's payment details which can be used by cyber criminals to defraud millions in a year.

According to a report, "World's retail-sales in e-commerce have reached 209% growth in revenue in a year.". The report further says that the rapid expansion in sales of e-commerce will further quantify. Transactions in Online mode have increased by 900% for Wine and crafts, this year.

Some reports also indicate that the departmental stores are expected to be shut by a huge margin of 60% whereas e-commerce is projected to grow by 20%. Thus, even the department retailers are expected to shift in e-commerce soon. Pandemic has approximately shifted the industry by 5 years.

As online business and retailers are reacting towards the fast development by finding better approaches to improve transportation, conveyance, & administration, there should be a solid emphasis on extortion location and security caused by online frauds. There are multiple online frauds like chargebacks, First party fraud, gaming and wireless fraud, Credit card fraud, etc.

In this paper, we specifically talk about credit card fraud. In online transactions credit cards are widely used as they save a lot of time for customers and the procedure of online payment is also very easy. This makes frauds of credit cards an easy target. Without much risk and easy hacking, a lot of money can be easily withdrawn from the credit card holder's account. The fraudsters try to show legitimate transactions which results in difficulty of detecting such fraudulent transactions.

We have used six classification models for predicting such fraudulent transactions. These models are-

1) *Decision Tree:* It is a machine learning algorithm in which we use labelled data and can be used for regression and classification problems. Decision Tree is mostly used in solving the classification issues.
2) *XG BOOST:* The algorithms are used and made for the performance and speed. It is a boosted version of the decision tree. It is an open-source library which can be used by everyone. It can run on Windows, Linux and macOS.
3) *K-nearest Neighbour:* K-nearest neighbours also known as KNN is also one of the supervised machine learning algorithms which is easy to implement and understand. These algorithms can solve the regression problems as well as classification problems.

4) *Logistic regression:* As Logistic regression uses true labels to train the model, it is supervised in nature and it can be used for solving classification problems. Here, Logistic regression calculation ought to have an objective variable (Y) and input factors (X) and when you train the model.
5) *Support Vector Machine:* Its a Linear classification model which is binary and has a decision boundary is constructed explicitly specially to minimize the errors is known as Support Vector Machine or popularly known as SVM. This is one of the machine learning algorithms which is very versatile and efficient.
6) *Random Forest:* The machine learning algorithm which uses the output of different decision trees combines to get a more precise and stable result which is very helpful in predicting the accurate result is known as Random Forest. It is a machine learning algorithm, supervised in nature which can be used for different types of tasks like regression and classification.

.

II. LITERATURE REVIEW

Extortion practices or false exercises are critical issues in numerous businesses like banking, protection, and so forth Particularly for the financial business, credit-card extortion is a major problem to determine. These enterprises endure a lot because of fake exercises towards income development and lose client's trust. So, these organizations need to discover extortion before it turns into a major issue for them.

Credit card exchanges for online instalments have expanded dramatically and misrepresentation endeavours on these instalments have gotten common with further developed hackers. Accordingly, regular extortion identification systems are not adequate to give sufficient precision to fraud recognitions. Machine learning classification models may give a proactive instrument to forestall credit card extortion with adequate precision.

Traditional methods for identifying frauds cannot identify complex fraudulent methods. To be restricted to an analysis of the card-holder’s behaviour, or to static guidelines of danger the executives of the frauds, had never halted the frauds to carry out their wrong-doings. Be that as it may, machine-learning models have had the option to address this issue, as we found in research. In some papers, users introduced a relative investigation of some machine learning procedures, which gave the best outcomes, as per their condition however applied to a similar dataset. The goal of the investigation was to pick the best credit card extortion detection procedures to actualize in future work.

Despite the fact that a huge number of studies have been done to tackle the issue of fraud detection, there is no commonly acknowledged solution or method. In many papers, openly accessible datasets were utilized. The unbalance issue of the data-set index was tackled by utilizing hybrid sampling techniques together. On this dataset index, comparative performance metrics evaluation have been performed. Not the same as different studies, the AUC (Area Under Curve) metric, which communicates the achievement in such datasets, has additionally been utilized notwithstanding standard evaluation metrics.

Some researchers also used machine learning methodology like SVM using spark to assemble prototypes addressing ordinary as well as strange client behaviour then it can be used for assessing the legitimacy of the latest transactions. Outcomes acquired by the data sets of Visa exchanges have shown that these strategies have become viable in the battle in case of banking frauds in big-data. Research results from one of the investigations have also shown that using SVM using Spark for prediction is preferable for evaluation.

III. METHODOLOGY AND OBJECTIVE

A. *Experimental Procedure*

The experimental procedure has the following steps which should be followed frequently. First of all, a dataset should be collected from which prediction should be done. Then after the dataset is further divided into testing and training of 10% and 90% respectively based on the requirements. Then the next step focus is on selecting an accurate model as per the requirements. For prediction purpose focus is on supervised machine learning models i.e., SVM, Logistic Regression, Decision Tree, XGBOOST, Random Forest and KNN. After this, the obtained testing set is entered into the models before examining the classification accuracy, F1 Score and Confusion Matrix. Moreover, the classifier’s training time is also evaluated to analyse the final complexities from training all the classifiers. Furthermore, finally check the accuracy, F1 score and confusion matrix for predicting the dataset or the model is achieving the goal as per the requirement or not.

B. Experimental Setup

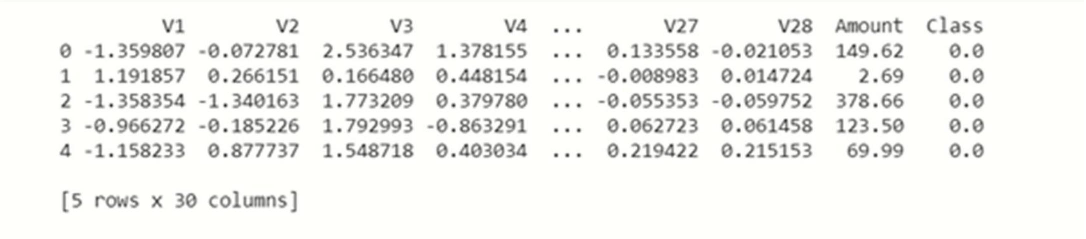
In this paper a dataset called creditcard.csv is used for all the machine learning models to predict the fraud detections in the credit card. This dataset has the purchases made by Credit cards for the year 2019. It shows that out of approximately two hundred eighty-four thousand transactions, there were a total four hundred ninety-two fraud cases. It is very uneven, there are only 0.172% cheat cases. Further we have divided the dataset into testing and training data. At last, the training datasets are fed into the model to predict the credit card fraud detection.

IV. INFORMATION ABOUT DATASET

Here, we use a dataset called creditcard.csv which is obtained from Kaggle, a website where datasets are available.

This dataset contains the exchanges that happened in only two days, in which we have 492 fakes out of 284,807 exchanges. This dataset is uneven because it contains positive classes or we can say that cheats represent 0.172%, here everything being equal.

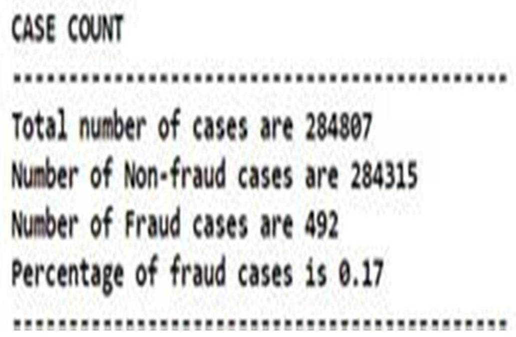
Because of the PCA change, it just contains mathematical information factors. This dataset contains such features i.e., V1, V2, V3, V4, …. V28 are the most important portions which are obtained from the PCA, there are also those features which are not changed with respect to PCA are known as “Time” and “Amount/Sum”. Here, the feature named ‘Time’ is used for the time taken between every trade and the primary trade for the dataset which is in seconds. Another feature named “Amount/Sum” is the exchange amount, this is used for instance dependent cost sensitive learning. And the last feature named “Class” is the



```
         V1        V2        V3        V4  ...       V27       V28  Amount  Class
0 -1.359807 -0.072781  2.536347  1.378155  ...  0.133558 -0.021053  149.62    0.0
1  1.191857  0.266151  0.166480  0.448154  ... -0.008983  0.014724    2.69    0.0
2 -1.358354 -1.340163  1.773209  0.379780  ... -0.055353 -0.059752  378.66    0.0
3 -0.966272 -0.185226  1.792993 -0.863291  ...  0.062723  0.061458  123.50    0.0
4 -1.158233  0.877737  1.548718  0.403034  ...  0.219422  0.215153   69.99    0.0

[5 rows x 30 columns]
```

Here, we first of all find out the case count then divide it into count for Non-fraud cases and Fraud cases. And, after that calculate the percentage of Fraud Cases.



```
CASE COUNT
********************************************
Total number of cases are 284807
Number of Non-fraud cases are 284315
Number of Fraud cases are 492
Percentage of fraud cases is 0.17
********************************************
```

Then, we found stats for each case (Fraud cases and Non-Fraud cases). Furthermore, we divided the dataset into testing data and training data. At last, we achieved the accuracy, F1 Score and Confusion Matrix for each model which are Decision tree, XGBOOST, k-nearest neighbour, Logistic-regression, Random forest and SVM. Were,

1) *Accuracy:* Accuracy is one measurement for assessing classification models. Casually, accuracy is the negligible portion of expectations our model got right. Officially, accuracy has the accompanying definition: Accuracy = Number of right predictions all out number of predictions;
2) *F1-Score:* The F1 score is also called the F-measure and F-score which is a proportion of a test's correctness. The F1 measure is characterized as the mean of the consonant(weighted) of the test's correctness as well as the recall.
3) *Confusion Matrix:* It is a tabular representation which is used for calculating the correctness or performance of the classifier or classification model for a given test dataset, of which know the truth values. Its representation is very simple and easy to understand.

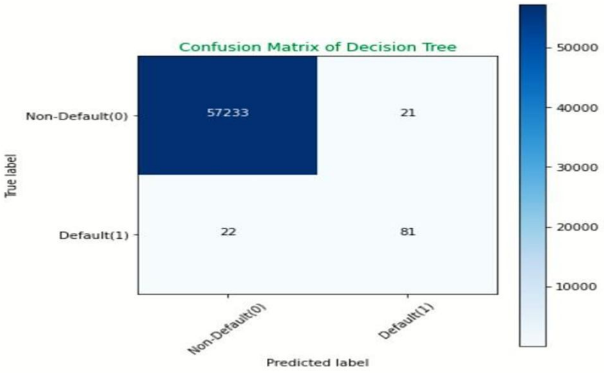
V. DECISION TREE

It is a machine learning algorithm in which we use labelled data and can be used for regression and classification problems. Decision Tree is mostly used in solving the classification issues. Decision Trees consist of nodes or leaves which are used for taking any decision and it also has different branches which helps the leaf nodes to make decisions, or we can say that leaf nodes are dependent on the branches and the most important thing is that leaf nodes have no further branches. Basically, the decision tree algorithm is a graphical representation for solving the issues or conditions which are dependent on the decisions. This algorithm is also known as a decision tree on lands because like a tree, it starts with a root node or hub which classifies or develops further different branches and at last it contains leaf nodes which do not contain any further branches as leaf on trees and that is builds or constructs like a tree like construction or structure.

The decision tree is the least complex and most famous order calculation. For building the model the decision tree algorithm considers all the given highlights of the information and thinks of the significant highlights. As a result of this preferred position, the decision tree algorithm likewise utilized in distinguishing the significance of the component measurements. Which is utilized in handpicking the highlights. When the significant highlights are distinguished then the model trains with the preparation information to concoct a bunch of rules. These standards are utilized in anticipating future cases or for the test dataset. Here, we will utilize the Decision Tree Classifier class from the sklearn library to prepare and assess models. We use X_train and y_train information for preparing purposes. X_train is a preparation dataset with highlights, and y_train is the objective mark.

A. *After Applying Decision Tree*

Confusion Matrix After Applying The Model



1) Accuracy After Applying The Model: 0.9992
2) F1 Score After Applying The Model: 0.79

VI. K- NEAREST NEIGHBOR

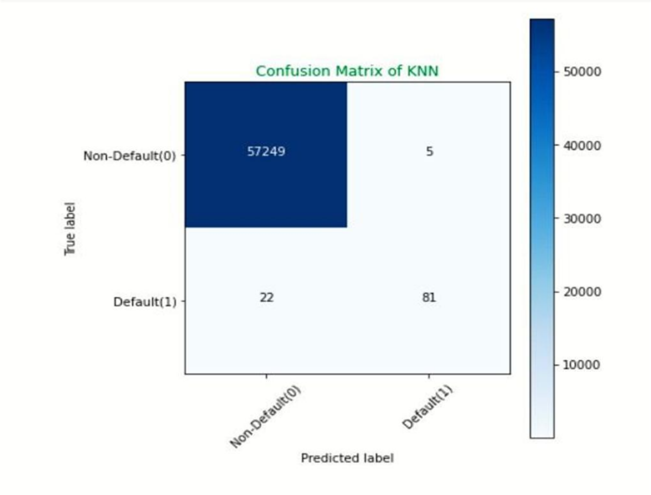
K-nearest neighbours also known as KNN is also one of the supervised machine learning algorithms which is easy to implement and understand. These algorithms can solve the regression problems as well as classification problems. It is supervised as you are attempting to order a point dependent on the known classification of different points. KNN algorithms use data and order new data points dependent on similitude measures (for example distance function).

Classification is finished by a dominant party vote to its neighbours. Its aim is to utilize a data in which data points are then isolated into a few classes to anticipate the grouping of another sample point. The KNN model can compete with the most accurate model because it gives highly accurate predictions so here, we use it for predicting the credit card fraud detection and we also use it because we do not require a human readable model. A calculation was made utilizing KNN which will examine a dataset to foresee fraud detection utilizing predefined attributes.

Since the KNN calculation requires no preparation prior to making expectations, new information can be added flawlessly which will not affect the exactness of the calculation so for accurate prediction of the model we use the KNN machine learning algorithm. Here, we also use some of the concepts of Random Forest and Bagging Classifier. Random Forest constructs are different, or we can say that multiple decision trees and then their outputs combine to get a more precise and stable result which is very helpful in predicting the accurate result of credit card fraud detection.

A. After Applying K Nearest Neighbor

Confusion Matrix After Applying The Model



1) Accuracy after applying the model: 0.9995
2) F1 score after applying the model: 0.85

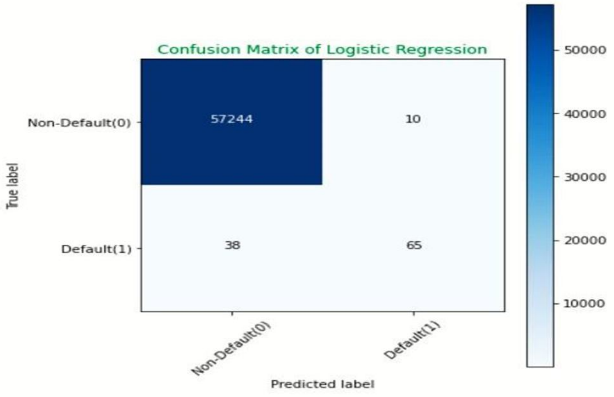
VII. LOGISTIC REGRESSION

It uses true labels to train the model and it can be used for solving classification problems; it is a supervised machine learning algorithm. Here, Logistic regression calculation ought to have an objective variable (Y) and input factors (x) and when you train the model. Logistic Regression is utilized to depict information and to clarify the connection between the dependent variable and independent variable. It is utilized to anticipate a parallel result dependent on a set of independent variables. Basically, it is utilized to figure the likelihood of a binary event happening, and to manage issues of classification and here we use it for predicting the fraud detection in credit cards.

By anticipating such results, calculated logistic regression causes data-analysts to settle on informed choices. Logistic Regression is a much easier method to train the model and implement it as compared to any other methods. In predicting fraud detection in credit cards, the dataset is linearly separable so here logistic regression works well.

A. After Applying Logistic Regression

Confusion Matrix After Applying The Model



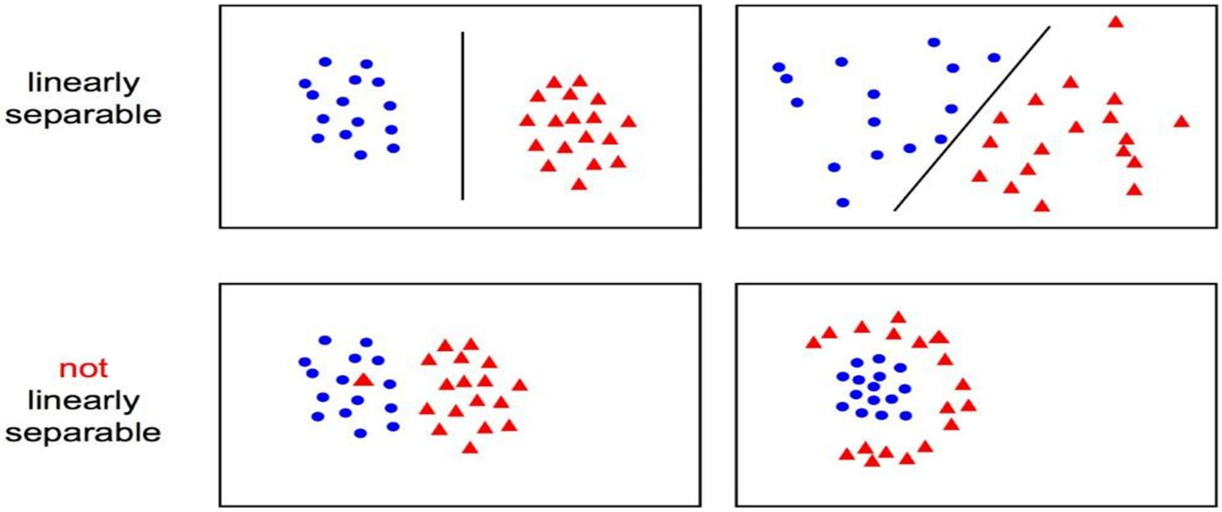
1) Accuracy after applying the model: 0.9991
2) F1 score after applying the model: 0.73

VIII. SUPPORT VECTOR MACHINE (SVM)

Its Linear classification model which is binary and has a decision boundary is constructed explicitly specially to minimize the errors is known as Support Vector Machine or popularly known as SVM.

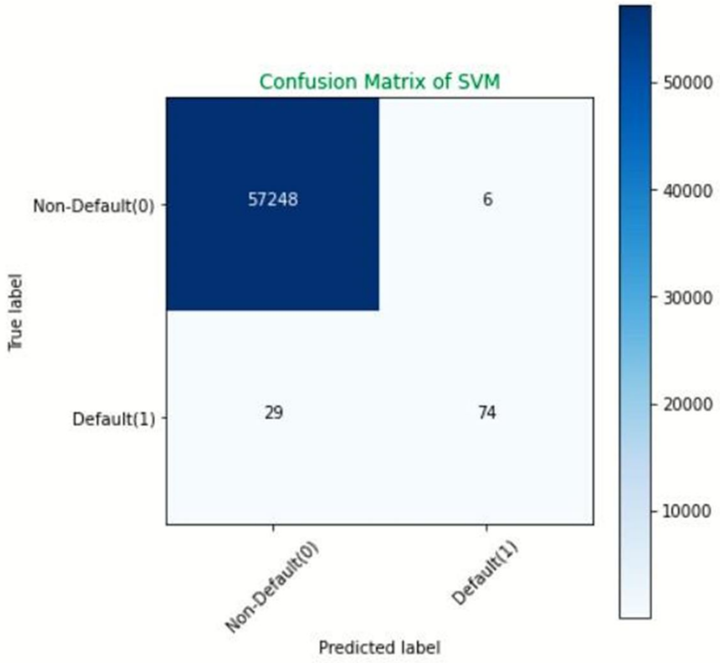
This is one of the machine learning algorithms which is very versatile and efficient can perform different types of tasks which includes regression, both the classifications which are linear and nonlinear.

It can also be used in outlier detection. Hyperplane which is used to classify the model is used by SVM for classification.



A. *After Applying The Model*

Confusion Matrix after Applying Svm



1) Accuracy after applying the model: 0.9993
2) F1 score after applying the model: 0.80

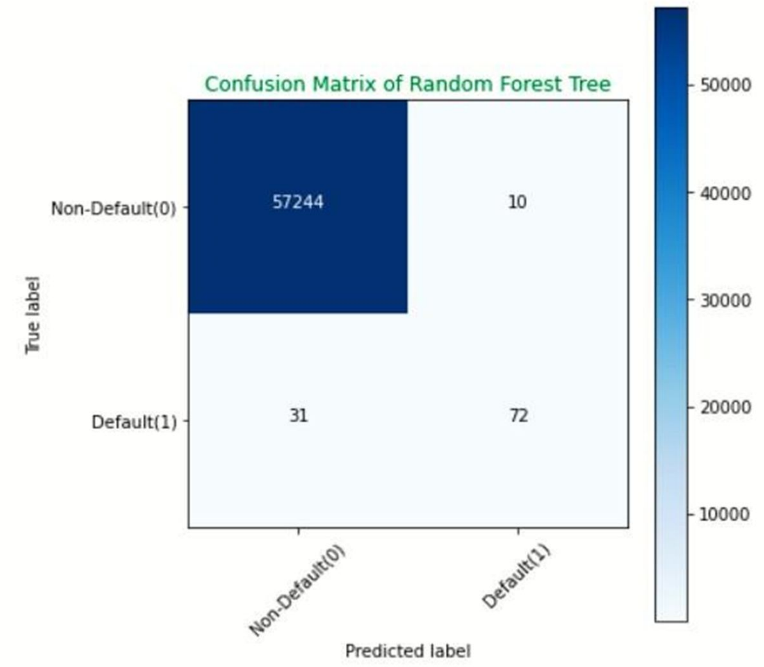
IX. RANDOM FOREST

The machine learning algorithm which uses the output of different decision trees combines to get a more precise and stable result which is very helpful in predicting the accurate result is known as Random Forest. It is a machine learning algorithm, supervised in nature which can be used for different types of tasks like regression and classification.

It can be constructed from different decision trees. It generally uses different models of decision trees for better accuracy and performance. These are one of the algorithms which are flexible and are easy to use. Random Forests are popular and widely used because of their diversity and simplicity. There are many advantages of Random Forests but the main among them is that it can be used for classification and also for the regression tasks. Here in this research paper, we have implemented the same algorithm for fraud detection in credit cards.

A. After Applying Random Forest

Confusion Matrix After Applying The Model



1) Accuracy after applying the model: 0.9992
2) F1 score after applying the model: 0.77

X. XG BOOST

The algorithms are used and made for the performance and speed. It is a boosted version of the decision tree. It is an open-source library which can be used by everyone. It can run on Windows, Linux and macOS.

A. Parameters

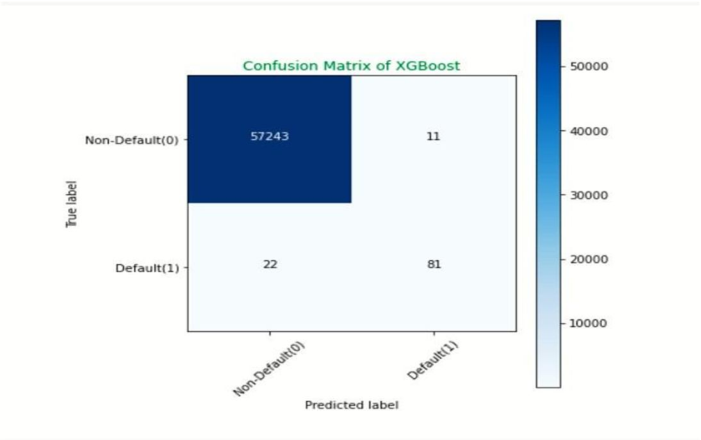
1) Silent (0 by default): if you want to print the running messages you need to specify 0 You need to specify 0 and for the silent mode use 1.
2) Booster: Its default value is gbtree. For using gbtree or gnlinear you need to specify the booster.
3) Num_pbuffer: default set by algorithm, there is no need for the user to set it.
4) Num_feature: default set by algorithm, there is no need for the user to set it.

Features of XGBoost algorithm includes:

a) XG Boost Does the Penalization of trees.
b) XG Boost shrink the leaf nodes.
c) XGBoost do the Newton Boosting.
d) XG Boost has additional randomization parameters .
e) XG Boost implements on single and distributed systems .

B. After Applying XG Boost

Confusion Matrix After Applying The Model



1) Accuracy after applying the model: 0.9992
2) F1 score after applying the model: 0.83

XI. RESULT

Machine Learning Algorithm Used	Accuracy	F1 Score
Decision tree	0.9992	0.79
K nearest neighbor	0.9995	0.85
Logistic regression	0.9991	0.73
Support Vector Machine	0.9993	0.80
Random forest	0.9992	0.77
XG boost	0.9994	0.83

XII. CONCLUSION AND FUTURE SCOPE

Credit card extortion is indeed a huge problem in this era. This paper has drilled down the most well-known strategies for detecting fraudulent alongside their identification techniques and checked on ongoing discoveries in this field. This paper has additionally clarified in detail, how various classification models can be used to improve detection of fraud alongside the algorithm, clarified its execution and experimentation results.

In this paper we have covered the in-depth research for Fraud Detection in Credit cards with the help of ML algorithms. We have predicted the frauds in credit cards using six machine learning models which are Decision tree, XGBOOST, k-nearest neighbour, Logistic-regression, Random forest and SVM. At last, we achieved the accuracy scores and f1 scores. Now looking at the high accuracies of each model, which is almost similar for all, we can say that all these six methods can be used for detecting frauds. But since technologies are continuously evolving, we can hope to see better accuracy for fraud detection in near future.

REFERENCES

[1] Roy, Abhimanyu, et al. "Deep Learning Detecting Fraud in Credit Card Transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS), 2018, doi:10.1109/sieds.2018.8374722.

[2] Xuan, Shiyang, et al. "Random Forest for Credit Card Fraud Detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, doi:10.1109/icnsc.2018.8361343.

[3] Awoyemi, John O., et al. "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative

[4] Jiang, Changjun et al. "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." IEEE Internet of Things Journal 5 (2018): 3637-3647.

[5] Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on AutoEncoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[6] Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[7] Randhawa, Kuldeep, et al. "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420. Analysis." 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, doi:10.1109/iccni.2017.8123782.

[8] Melo-Acosta, German E., et al. "Fraud Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques." 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, doi:10.1109/colcomcon.2017.8088206.

[9] 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.

[10] 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.

[11] 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.

[12] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai, S. Pan, "Credit card fraud detection based on whale algorithm optimized BP neural network", 2018 13th International Conference on Computer Science & Education (ICCSE) pp. 1-4. IEEE

[13] 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.

[14] 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

[15] Learning – Towards Data Science. [online] Available at: https://towardsdatascience.com/deeplearning-vs-classical-machine-learning-9a42c6d48aa [Accessed 19 Jan. 2019].

[16] Kaggle.com. (2019). Credit Card Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud [Accessed 10 Jan. 2019].