Credit Card Fraud Detection Using Machine Learning

T.C.Ezhil Selvan[1], A.Nithya[2], R.Sandhiya[3], T.Devadharshini[4]
[1]Associate Professor, Dept of IT
[2, 3, 4]Dept of IT
[1, 2, 3, 4] Sri Ramakrishna Institute Of Technology

Abstract- *The development of technologies like e-commerce and financial technology apps has led to an increase in the daily volume of online card transactions. The ability to misuse credit cards has also increased. The outcome is an increase in credit card fraud that impacts banks, merchants. The Financial institutions as well as cardholders suffer considerable financial losses as a result of credit card theft. Therefore, developing systems to ensure the security and precision of credit card transactions is essential. Using real-world datasets compiled from European credit cardholders construct a Machine Learning (ML) based framework for credit card fraud detection in this study. The primary goal of this project is to identify such credit card frauds using few Machine Learning techniques such as Decision Tree, Random Forest, Logistic Regression, and XG Boost. The prediction of different algorithms can be made with more accuracy by comparing the aforementioned algorithms*

Keywords- Accuracy, Credit Card Fraud Detection, Decision Tree, Imbalanced dataset, Logistic Regression, Random Forest.

I. INTRODUCTION

In this paper, Machine Learning (ML) algorithms for credit card fraud detection that are evaluated on a real-world dataset which was generated from European cardholders in September 2013. This dataset is highly imbalanced. To alleviate the issue of class imbalance that is found in the European card dataset, this research investigated the use of the Synthetic Minority Over-Sampling Technique (SMOTE). In recent years credit card transaction fraud is a major issue in various sectors like banking, insurance, finance, etc. Fraud is a set of illicit activities to obtain goods and funds. In the traditional approach, the algorithm, were written based on strict rules. If a new fraud is detected, then new changes in algorithm is done by fraud analyst either by writing new algorithm or by changing the already existing algorithm. All these changes are done by fraud analyst. In this approach, as number of customers and data increases, human efforts also increase. Also, in the real world FDS, investigator is not able to check all transactions alert. In this approach, the Fraud Detection System monitors all the approved transactions and it alerts the most doubtful one. Verification of all these suspicious alerts is done by investigator and feedback is provided to FDS which indicate that if the transaction was authorized or fraudulent. So, this traditional rule-based approach is time-consuming and costly. Only few alerts each day are verified while the rest of the transactions alert remains unchecked until customer identifies them and report them as a fraud. Also, fraudsters use diverse techniques in finding a loophole in FDS so that they can do their illegal activities in transaction. So, most of the time it is hard to identify fraud in the credit card transaction. The reason of such illegal transaction might be to get items without giving money. Identifying such illegal activities or fraud is a troublesome and may risk the business and business organizations. Moreover, the ML methods that were considered in this research include: Random Forest (RF), Extreme Gradient Boosting (XG Boost), Logistic Regression (LR), and Decision Tree. All ML algorithm is evaluated using a real-world credit card transaction. The main motive of this paper to apply supervised learning method on the real- world dataset to identity fraud or nonfraud transaction.

II. PROBLEM STATEMENT

In today’s world, lot of fraudulent activities are happening. Credit Card Frauds are increased at a high rate. The payment process may be done either virtual or credit card. The physical card payment means payment by swiping the card in the machine. The virtual card payments are just online payment where its fraud rates are very high. The CCF detection system is done manually. The existing system is extravagant and time consuming, it involves lot of paperwork. To overcome such problems, classification algorithms in Machine Learning were used. More than 7 algorithms were compared to predict best accuracy associated with deep learning. In this project, the four supervised learning classification algorithms in Machine Learning are used to detect these fraud transactions in credit card fraud detection.

III. RELATED WORK

Fawaz Khaled Alarfaj et al., proposed the method, CCF is becoming a bigger threat to financial institutions [3]. New fraud methods are regularly developed by fraudsters. The changing fraud landscape can be managed by a powerful classifier. The main goal of a fraud detection system is to reduce the number of false-positive cases while precisely anticipating fraud events. Depending on the particular business environment, ML techniques operate differently. The type of input data heavily influences the various ML strategies. The number of characteristics, the volume of transactions, and the correlation between the features all have a significant impact on how well the model detects CCF. Deep Learning techniques, such as CNNs and their layers, are connected to text processing and the baseline model. When the performances of each algorithm are compared head-to-head, the baseline model and the CNN with 20 layers come out on top with a 99.72% accuracy rate.

Altyeb Altaher Taha and Sharaf Jameel Malebary proposed the method for identifying fraud in credit card transactions, employs an Optimized Light Gradient Boosting Machine (OLightGBM) [1]. In two real-world data sets, performed a number of tests. State-of-the-art machine learning algorithms, including random forest, logistic regression, the radial support vector machine, the linear support vector machine, k-nearest neighbours, decision tree, and naive bayes, were used to compare the performance of the proposed approach with other research findings and results. The experimental outcomes show that the suggested strategy beat the competing machine learning algorithms and attained the maximum accuracy, Area under curve, precision, and the score values. The findings show that the suggested method outperforms alternative classifiers [4]. The outcomes also demonstrate the significance and benefit of implementing an effective parameter optimization technique for boosting the predictive effectiveness of the suggested approach.

Emmanuel Ileberi et al.., proposed, using the credit card fraud dataset, this article constructed a number of ML algorithms for credit card fraud detection. The DT, RF, ET, XGB, LR, and SVM are the ML techniques that were suggested in this paper [2],[5]. Each of the suggested techniques was also combined with the AdaBoost method to improve classification accuracy and address the problem of class imbalance. Additionally, a comparison analysis between the techniques in this work and current framework for detecting credit card fraud was done. For instance, 99.67%, 99.95%, 99.98%, and 99.98% accuracy were attained by the DT-AdaBoost, RF-AdaBoost, ET-AdaBoost, and XGB-AdaBoost, respectively. The XGB-AdaBoost and ET-AdaBoost both received MCCs of 0.99 for the quality of their categorization results. These results proved that applying the AdaBoost algorithm improves the suggested ML techniques. A highly skewed synthetic credit card fraud dataset was used to validate the methodology suggested in this study, and the outcomes were excellent. The ET-AdaBoost, for instance, the achieved an accuracy of 99.99% and an MCC of 0.99. Additionally, the AUC value of 1 was obtained by the XGB-AdaBoost, DT-AdaBoost, ET-AdaBoost, and RF-AdaBoost. In future works, intend to test and validate the proposed framework on additional credit card fraud datasets that will be sourced from financial institutions.

V. N. Dornadula and S. Geetha is proposed the E-commerce and many other online websites have enlarged the online payment modes, growing the risk for online frauds [7][6]. Increase in fraud rates, researchers started using diverse machine learning methods to detect and analyze frauds in online transactions. The main aim of the paper is to plan and evolve a novel fraud detection method for Streaming Transaction Data, with an intent, to analyze the past transaction details of the customers and extract the behavioural figured and where cardholders are clustered into different groups based on their transaction amount. After that using sliding window strategy, to aggregate the transaction made by the cardholders from various groups so that the behavioural pattern of the groups can be withdraw. Later different classifiers are trained over the clusters separately, then the classifier with finer rating score can be chosen to be one of the best methods to predict frauds. This worked with European credit card fraud dataset and also using Support Vector Machine (SVM) to obtain better accuracy.

IV. PROPOSED SYSTEM

In this work, fraud detection using machine learning is proposed was designed a model to detect the fraud activity. The classifier used by the proposed method, to minimize credit card fraud is built using classification of algorithm like Logistic Regression, Decision Tree, Random Forest, XG Boost for better performances. Adasyn, SMOTE, Undersmapling technique are paired with these algorithms to ensure a high degree of detection accuracy, recall, Receiver Operating Characteristic (ROC). The aim of the credit card fraud detection project is to predict whether a transaction is fraud or not based on the transaction location, age, gender etc. in the dataset.

V. ALGORITHMS

A. Logistic Regression

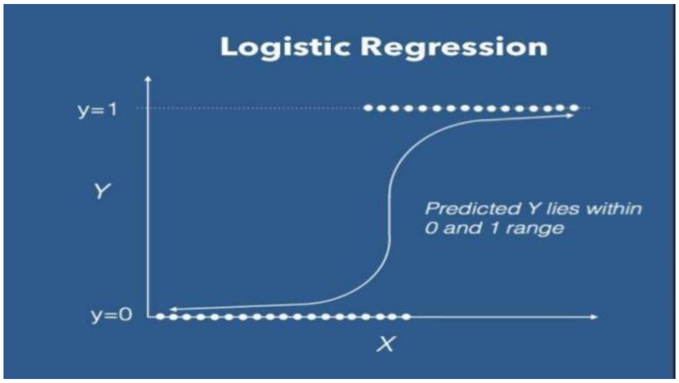
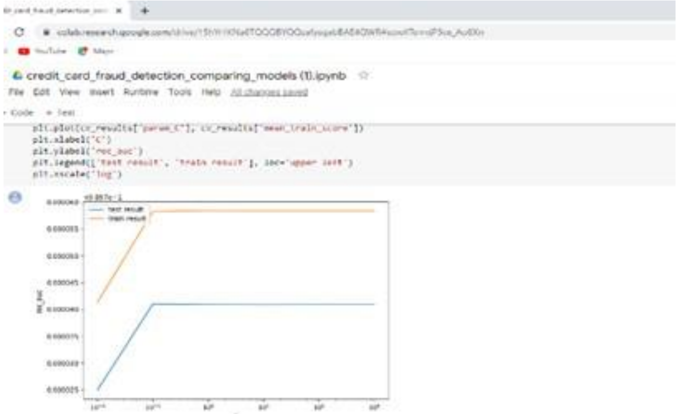


Fig 1: Logistic Regression

The classification and output of Logistic Regression (LR) are discrete values It belongs to the Supervised Learning approach and is one of the most widely used Machine Learning algorithm. It provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc. In this algorithm, with the help of dataset, can begin the process of training the logistic algorithm. It passes input data to the algorithm and using the known outcomes, train the algorithm to predict the correct result. To test the accuracy of algorithm, need to use separate dataset of input features and known outcomes. Then apply logistic algorithm to this testing dataset, and compare the predicted outcomes to the known outcomes. To attain better accuracy, the different techniques were used such as Over Sampling, Under Sampling, SMOTE and so on and these are combined with the algorithms, then involve changing the hyperparameters, ROC to improve performances. ROC is used for identifying the best model.


credit_card_fraud_detection_comparing_models (1).ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
test result
train result

```
plt.plot(cv_results['param_C'], cv_results['mean_train_score'])
plt.xlabel('C')
plt.ylabel('roc_auc')
plt.legend(['test result', 'train result'], loc='upper left')
plt.xscale('log')
```

Fig 2: Logistic Regression ACC

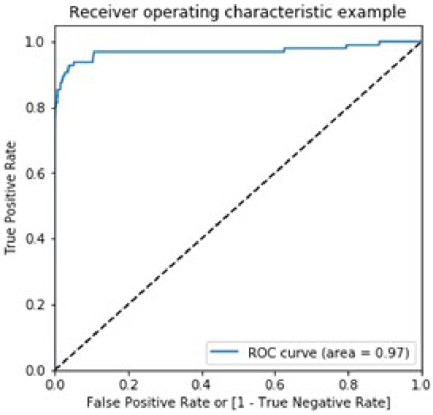


Fig 3: ROC score of LR with SMOTE technique

B. DECISION TREE

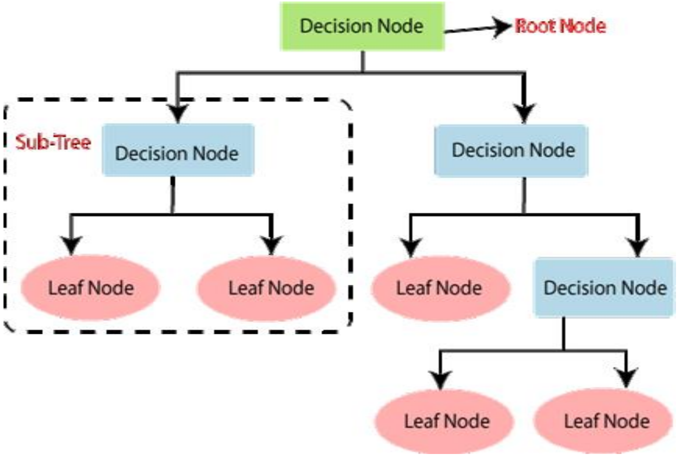


Fig 4: Decision Tree

The Decision Tree (DT) algorithm is a popular machine learning algorithm used for both classification and regression tasks. The algorithm starts with the entire dataset and selects the feature that splits the dataset in the most effective manner. The process is then repeated for each of the subsets until a stopping criterion is reached. This data should be in a format that is suitable for the decision tree algorithm (i.e. structured data). Once the data has been collected, it should be split into training and testing sets. The training set is used to train the decision tree algorithm, while the testing set is used to evaluate its performance. This involves selecting the appropriate hyper parameters for the model and fitting it to the training data. After training the model, it should be tested using the testing set. This involves evaluating the performance of the model using metrics such as accuracy, precision, recall, ROC and F1 score. If the model does not perform well on the testing data, adjustments can be made to improve its performance. This could involve changing the hyper parameter, adding or removing feature, or trying different algorithms.

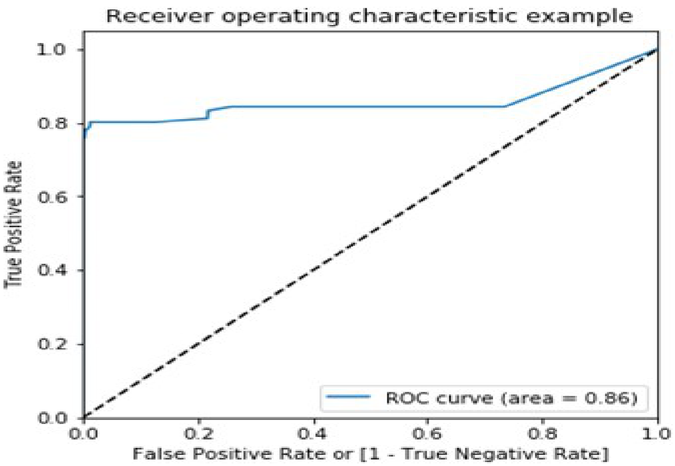


Fig 5: Decision Tree with SMOTE technique

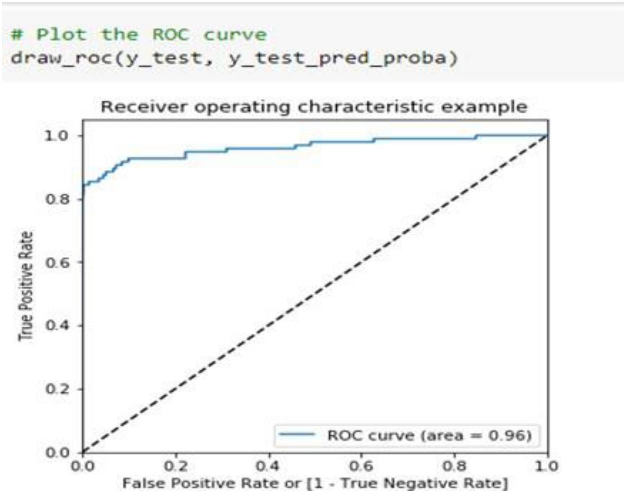
C. RANDOM FOREST

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It combines multiple decision trees to create a more accurate and robust model. To build a random forest model using the training data. This involves selecting a suitable number of decision trees, choosing hyperparameters such as the depth of the trees or the number of features to consider at each split, and fitting the model to the training data.

After building the random forest model, need to evaluate its performance using the test data. This can use various metrics such as accuracy, precision, recall, F1 score, or ROC curve to measure the performance of the model. It is also important to check for overfitting and tuning the hyperparameters to improve the model's performance.

D. XGBOOST

The Extreme Gradient Boosting (XGBoost) algorithm is a popular machine learning algorithm. The dataset taken from Kaggle can be used to train the XGBoost model for the purpose of credit card fraud detection. This algorithm is known for its efficient computation speed and improved performance compared to other machine learning algorithms. When applied to credit card fraud detection, the XGBoost algorithm can effectively identify fraudulent transactions by learning patterns and abnormal behavior in the dataset. This can be done by analyzing various factors such as transaction amount, merchant location, and transaction history. The model has to trained and tested using dataset to perform and then result will be predicted. This algorithm is a powerful tool for credit card fraud detection and can significantly improve the accuracy and effectiveness of fraud detection systems. The performance of the XGBoost algorithm in terms of its accuracy, precision, recall, and F1 score.



```
# Plot the ROC curve
draw_roc(y_test, y_test_pred_proba)
```

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.96)

Fig 6: ROC score of XG BOOST with SMOTE technique

VI. IMPLEMENTATION

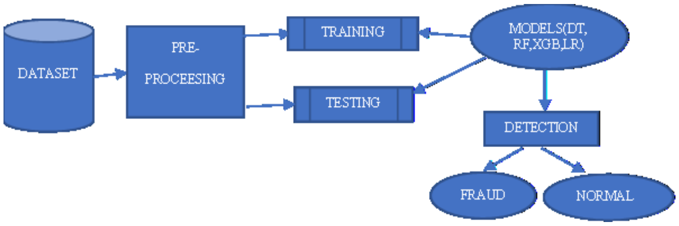


Fig 7: System Architecture

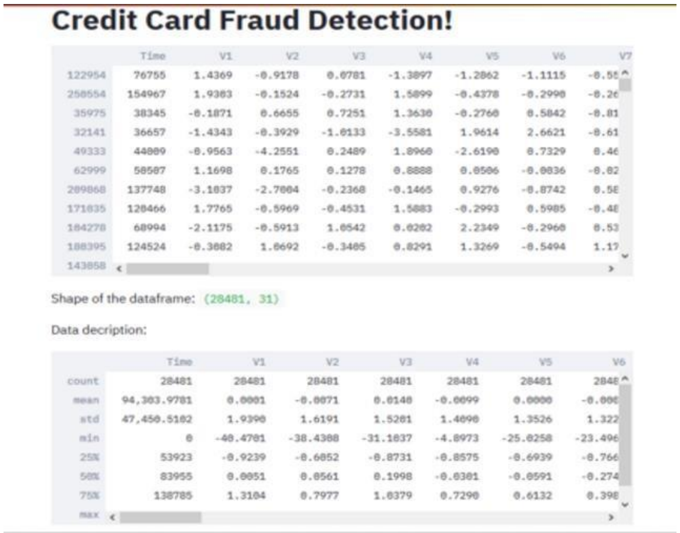
The software and hardware requirements used in this project includes;

TABLE I. TOOLS USED

Operating System	Windows 10 or above
Language	Python3.5
Editor	Google Colab, V Scode
Processor	corei3 or above
RAM	min 2GB

1. DATASET

A dataset is a collection of data that relates to a certain subject (contains diverse attributes). There are several techniques to collect the data, like web scraping, manual interventions, etc. The attributes are name, gender, age, transaction and so on. In this dataset, there are 21 attributes. The dataset consists of more than 400 individual data collected from internet. There are 30 columns in the dataset, which are described below,



Credit Card Fraud Detection!

	Time	V1	V2	V3	V4	V5	V6	V7
122954	76755	1.4369	-0.9178	0.0781	-1.3897	-1.2862	-1.1115	-0.55
258554	154967	1.9383	-0.1524	-0.2731	1.5899	-0.4378	-0.2998	-0.26
35975	38345	-0.1871	0.6655	0.7251	1.3630	-0.2760	0.5842	-0.81
32141	36657	-1.4343	-0.3929	-1.0133	-3.5581	1.9614	2.6621	-0.61
49333	44009	-0.9563	-4.2551	0.2489	1.8960	-2.6190	0.7329	0.46
62999	50507	1.1698	0.1765	0.1278	0.8888	0.0506	-0.0836	-0.82
209868	137748	-3.1837	-2.7004	-0.2368	-0.1465	0.9276	-0.8742	0.58
171035	120466	1.7765	-0.5969	-0.4531	1.5883	-0.2993	0.5985	-0.48
104278	68994	-2.1175	-0.5913	1.0542	0.0202	2.2349	-0.2960	0.53
188395	124524	-0.3082	1.0692	-0.3405	0.8291	1.3269	-0.5494	1.17
143858								

Shape of the dataframe: (28481, 31)

Data decription:

	Time	V1	V2	V3	V4	V5	V6
count	28481	28481	28481	28481	28481	28481	2848
mean	94,303.9781	0.0001	-0.0071	0.0148	-0.0099	0.0000	-0.006
std	47,450.5102	1.9390	1.6191	1.5201	1.4090	1.3526	1.322
min	0	-40.4701	-38.4388	-31.1037	-4.8973	-25.0258	-23.496
25%	53923	-0.9239	-0.6052	-0.8731	-0.8575	-0.6939	-0.766
50%	83955	0.0051	0.0561	0.1998	-0.0301	-0.0591	-0.274
75%	138785	1.3104	0.7977	1.0379	0.7290	0.6132	0.398
max							

Fig 8: Dataset

2. DATA PREPROCESSING

This includes cleaning, normalizing, and transforming the data to make it suitable for machine learning algorithms. In this phase, the data which are in the form of matrices, text format are pre-processed into numeric values, this can be done with the set of libraries.

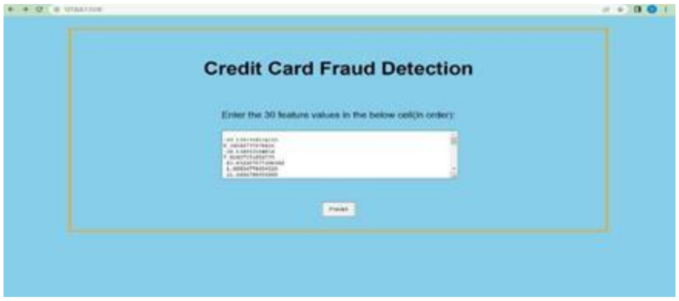
3. DATA SPLITTING

After pre-processing the data, can split it into training and testing sets. The train data is used to train algorithm, test data is used to evaluate model's performance.

4. MODEL SELECTION

A machine learning algorithm is selected based on the type of problem and the characteristics of the data. Divide the data into two sets with 80:20 proportion between training and testing. The machine learning algorithm is trained using the training set to learn from the data and make predictions or classifications and then testing set are used. Once the training and testing of the data is complete, save it using a library like a pickle into a ". pkl" file. Then predict the accuracy.

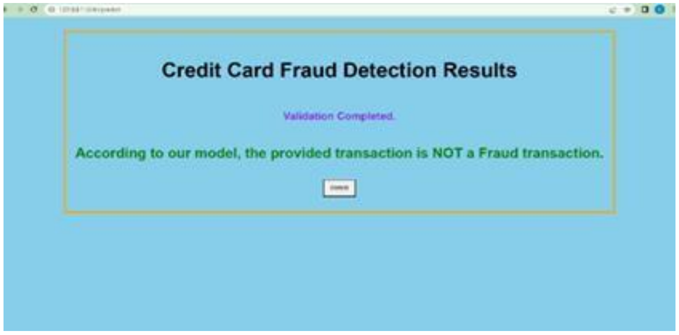
Web Page created to detect the given dataset transaction is fraud or non-fraud using the libraries like flask and appropriate language.



Credit Card Fraud Detection

Enter the 30 feature values in the below cell(in order):

Fig 9: Home page with imported data

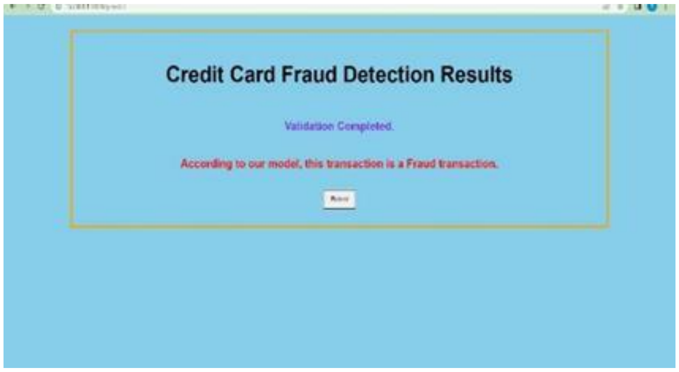


Credit Card Fraud Detection Results

Validation Completed.

According to our model, the provided transaction is NOT a Fraud transaction.

Fig 10: Detects Not Fraud



Credit Card Fraud Detection Results

Validation Completed.

According to our model, this transaction is a Fraud transaction.

Fig 11: Detects Fraud

VII. RESULTS

By comparing the classification of algorithms such as Logistic Regression, Decision Tree, Random Forest, XG Boost are determined to find and predict the best accuracy. Even though XG Boost, Random Forest performs well, the Logistic regression model is the best model to choose because of the easy interpretation of models and also the resource requirements to build model is lesser than other heavy models. After performing several models, the balanced dataset with SMOTE technique the simplest Logistic Regression model has good ROC score and also high recall. The ROC score (accuracy) of the algorithm with shows best model as outcome described below,

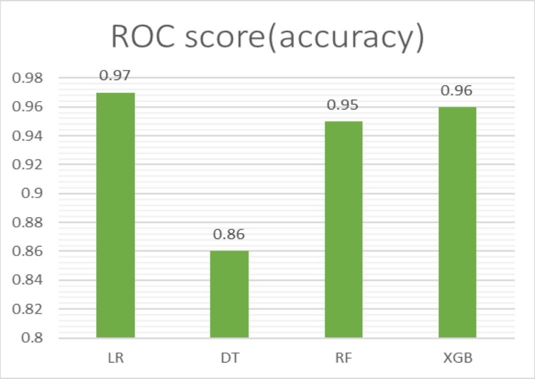


Fig 12: ROC score of the algorithms

VIII. CONCLUSION AND FUTURE WORK

Machine Learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XG Boost is used to detect the fraud in credit card system. In this project the supervised learning algorithm are used to predict the better accuracy. The different techniques are used to balance the imbalanced data, to improve the performances. With use of diverse parameters (i.e., Hyper-parameter tuning) the algorithms are performed. Comparing all algorithm performances, the logistic regression with SMOTE technique gives better ROC- accuracy of 97.0%. Future work, may explore unsupervised algorithms and neural concepts, large data to improve performance of the model proposed in this study. And also provide message alert for the customer if misuse happen by fraudster.

REFERENCES

[1] Altyeb Altaher Taha and Sharaf Jameel Malebary, “An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine”, IEEE Access vol.8, pp.25579 – 25587, 03 Feb. 2020.

[2] Emmanuel Ileberi et al.., “Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost” IEEE Access Volume: 9, pp.165286 – 165294, 15 Dec. 2021.

[3] Fawaz Khaled Alarfaj et al., “Credit Card Fraud Detection Using State-of-the- Art Machine Learning and Deep Learning Algorithms”, IEEE Access Volume: 10, p.39700 – 39715, 12 Apr.2022.

[4] HaiboWang; Wendy Wang; Yi Liu; Bahram Alidaee, “Integrating machine learning algorithms with quantum annealing solvers for online fraud detection”, IEEE Access., vol. 10, pp.75908 – 75917, 14 Jul.2022.

[5] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection", IEEE Access, vol. 7, pp. 93010-93022, 2019.

[6] S.Warghade, S. Desai and V. Patil, "Credit card fraud detection from imbalanced dataset using machine learning algorithm", Int. J. Comput. Trends Technol., vol. 68, no. 3, pp. 22-28, Mar. 2020.

[7] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms", Proc. Comput. Sci., vol. 165, pp. 631-641, Jan. 2019.

[8] Y. Lucas and J. Jurgovsky, "Credit card fraud detection using machine learning: A survey", arXiv:2010.06479, 2020.