Credit Card Fraud Detection Using Machine Learning

Sandeep Kumar
Associate Professor
Cambridge Institute Of Technology

Kavana Krishnappa
Student
Cambridge Institute Of Technology.

***Abstract*—Credit card is the commonly used payment mode in the recent years. As the technology is developing, the number of fraud cases are also increasing and finally poses the need to develop a fraud detection algorithm to accurately find and eradicate the fraudulent activities. This research work proposes different machine learning based classification algorithms such as logistic regression, random forest, and Naive Bayes for handling the heavily imbalanced dataset. Finally, this research work will calculate the accuracy, precision, recall, f1 score, confusion matrix, and Roc-auc score.**

***Keywords*—Fraud detection, Credit card, Machine learning, Accuracy, F1 score, Precision, Recall, Roc-auc score, Confusion matrix**

I. INTRODUCTION

The primary objective of this research work is to identify the fraudulent transactions using credit cards. To accomplish this, it is required to classify the fraudulent and non-fraudulent transactions. The primary goal is to make a fraud detection algorithm, which finds the fraud transactions with less time and high accuracy by using machine learning based classification algorithms. As technology is advancing rapidly, the payment by cash is reduced and online payment gets increased, this paves way for the fraudsters to make anonymous transactions.

In some modes of online payments, only card number, e xpiration date, and cvv are required and that data may be lost without our presence, in some cases we don't even know our data is being stolen.

The purchases that done over the internet where fraudsters use phishing techniques to grab the details still, we do not know that our data has leaked. To do fraud he just needs card details for some purchases and the user may not know whether his/her credit card information was leaked. The card details should be kept private. But sometimes it is not in our hands. Due to phishing sites the information may be leaked, Sometimes the card itself may be lost or may be stolen. The best way to find whether a t ransaction is fraud or not we need to find the spending pattern of the customer by using existing data and use Machine learning to find whether a is genuine or not.

A. Types of Frauds:

- Online and Offline
- Card Theft,
- Data phishing
- Application Fraud
- Telecommunication Fraud

II. BACKGROUND AND RELATED WORK

Fraud in any way is a criminal activity and is an offence, credit card fraud is stealing money. There are many studies in which they tried to find whether a transaction is fraud or not. Still having many challenges and tries to overcome those problems [7]

Firstly, many used Data Mining Techniques to find fraudulent transactions by using some Traditional approach, which is not conventional and these days fraudsters are so smart that they can do

fraud without violating rules [6]. so, Using Machine learning is conventional.

Machine learning also comes with challenges. So, here a heavily imbalanced data (Data set from Kaggle by mlg-ulb j) set is considered, so that it will give us the best algorithm to use along with its challenges. As it is heavy imbalanced, even if the proposed algorithm is good or not, it gives us an accuracy of about 99.9%. So, here the under sampling is considered to provide us with good results as in [8], Outlier detection and removal algorithms are used to accurately predict fraudulent transactions of a credit card transaction dataset as in [11].

Outlier data is used to deal with detecting the anomalous activities. Even after a huge number of proposed algorithms and mechanis ms to stop fraudulent transactions, the fraudsters are so clever that they always try to find new ways to make anonymous transactions and sometimes even the proposed algorithm could not find whether the transaction is fraud or not. So, to stop these frauds, the proposed algorithm should be made to learn fromthe past frauds and use it for future frauds, which can even detect the fraud before it takes place.

III. PROPOSED METHODOLOGY

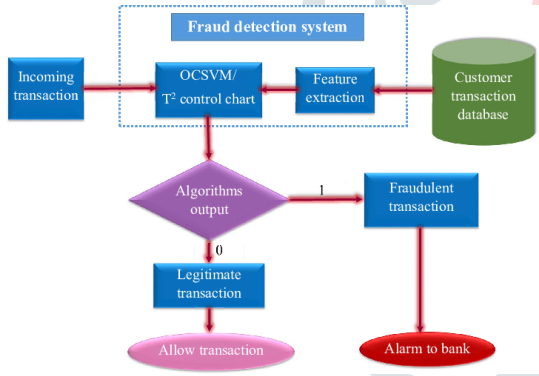

Fig. 1.0 Workflow overview. (Figure 1.0 depicts workflow and steps of our implementation to reach the objective.)

The dataset contains transactions of European credit card holders of September 2013 for two days it contains v1-v28 PCA [4] feature because of confidentiality issues and Time, Amount, which are known features and class with 0 and 1 where 1 means fraud 0 means non fraud.

Table 1- Dataset

	V1	V2	V3	V4	V5
0	-1.359807	-0.072781	2.536347	1.378155	-0.338321
1	1.191857	0.266151	0.166480	0.448154	0.060018
2	-1.358354	-1.340163	1.773209	0.379780	-0.503198
3	-0.966272	-0.185226	1.792993	-0.863291	-0.010309
4	-1.158233	0.877737	1.548718	0.403034	-0.407193
5	-0.425966	0.960523	1.141109	-0.168252	0.420987

We can see from Table 1 sample of the dataset and few features v1-v5.

Fig. 1.1 Visualization of Data Distribution. (As fig1.1 shows it is heavily imbalanced.)

After performing Under Sampling, the dataset should be reduced to 890 from 284,807 transactions.

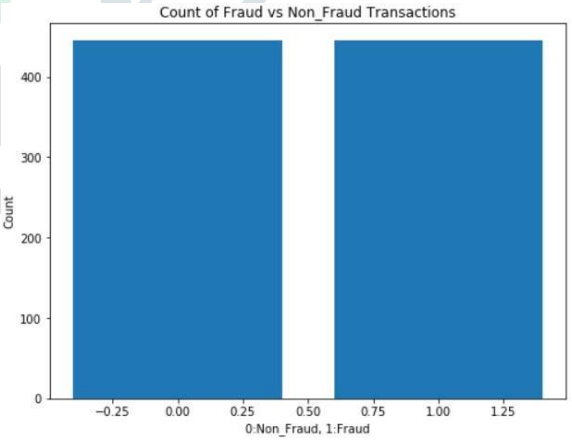

Fig. 1.2 Describes visualization of data distribution after under sampling. (As fig1.2 shows the data is balanced.)

As discussed using the Outlie r Data min ing technique, a better accuracy can be obtained. Because the bias and extreme values are identified and

removed, better results can be obtained so now our task is to detect the extre me outliers and remove them. First, a heat map will be obtained for visualizing the correlation between variables.

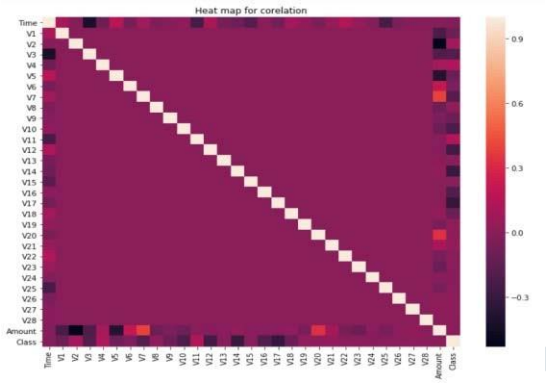

Fig. 1.3 Describes heat map for finding correlation. (As fig1.3 shows all attributes are correlated.)

The class column is removed, and time along with the amount of columns are also normalized. By removing class columns from the dataset, the evaluation will be fair and true positive rate also increases. The dataset is now formatted and processed. By finding features with high positive correlation and high negative correlation, those features will be removed in order to increase the accuracy [3].

Removed features are.

Table 2

	Class
V3	-0.562
V4	-0.564
V5	-0.628
V6	-0.685
V7	-0.752
V8	-0.599
V9	-0.555
V4	0.716
V11	0.684

The above table show us the features having high correlation, which can make the proposed algorithm more biased and may affect our output. Not only that, the values can also be scaled in a range, which in turn results in good outcome.

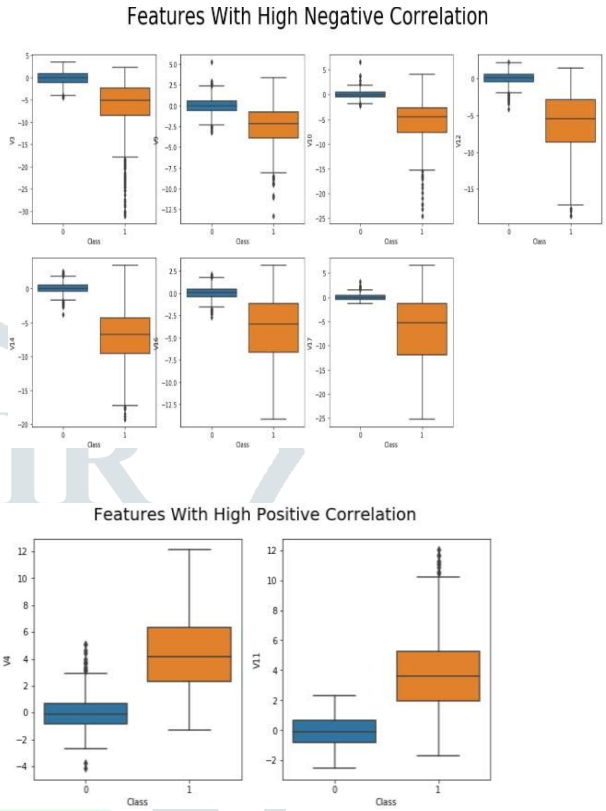

Fig. 1.4,1.5- Describes features having high correlations.

So, after outlier detection and removal we again decreased to 620 transactions from 890.

The dataset is divided, 70% for training 30% for testing after doing preprocessing techniques discussed. Now we are going to imple ment our algorithms.

Decision Tree: A Decision tree is a tree like structure where leaves are decisions made from nodes which contains data and used for classification and prediction, where variables are trained to predict output here these trained variables form as branch and decides output and again these branches form as tree and finally output is predicted.

ID3 (Iterative Dichotomiser) fo r informat ion gain this is used to select the branch which gives more gain that will be selected. Here we will decide the root node by calculating information gain using entropy.

$$info\,(D) = -\sum_{i=1}^{m} plog2\ \pi\iota \qquad (1)$$

Eq. 1 ID3 algorithm formula, for calculating Information Gain.

The equation mentioned above is used for Decision tree in order to make it as root node by knowing the information gain. -eq (1)

Random Forest: Random Forest is a classification algorithm in which it contains many numbers of decision trees for different subsets of the dataset and average of all decision trees accuracy and improves the total accuracy.
If number of decision trees increases the accuracy of random forest also increases
The random forest is same as decision tree, but it contains man of them from which a better outcome can be expected.

Logistic Regression**:** It is used for both classification and regression, but it is widely used for classification. The output is a binary belonging to one of the classes. It is used to predict output with the help of dependent variables. This algorithm easily binary classification to two values 0 or 1

$$p = 1/1 + e^{-(a0+a1x1+a2x2+\cdots\ldots+anxn)} \quad (2)$$

Eq. 2 Explains the principle and how the logistic Regression works.
In above equation a0+a1x1+a 2x2+⋯ +an xn a0, a1…. an. are coefficients and x1, x2…...xn are independent variables, p is outcome.

Naive Bayes: A classification algorithm which uses Bayesian principle to find the output.
It takes the probability of an event (feature) with that probability it calculates the probability of another output and based on it shoes whether it is fraudulent transaction or not.

$$p(c/x) = p(\mathrm{X}\ 1/c) * p(x\,2/c) *.. \\ * p(xn/c) * p(c) \quad (3)$$

Eq. 3 Explains the principle of Naive Bayes and how it works. The above equation is outcome of probability events naive bias uses Bayesian principle for predicting the outcome.

Confusion matrix: The confusion matrix provides more knowledge about the performance of our model by providing the information of correctly, incorrectly classified classes through which we can identify errors.

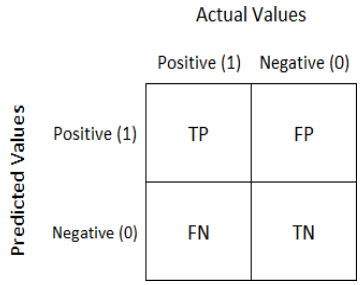

Fig 1.2 - As Shown how confusion matrix forms.

Accuracy: Accuracy is the percentage of correctly predicted outputs

Precision: number of classified correct outputs we can say exactness of model.

Recall: the measure of our model correctly identifying True Positives

F1 score: Average of Precision and Recall.

IV. EXPERIMENTATION AND RESULTS

Here I measured accuracy, recall and precision score, f1 score for every algorithm ROU-AUC score and Confusion Matrix for finding the best algorithm.

Now we will show the results we got after doing data mining techniques as we need to get knowledge about which algorithm works perfectly. (class 0 means Genuine transaction and 1 means fraud transaction.)

Here to show why we need to use data mining technique we trained data set without using data mining with 'Decision Tree classifier' and we got an accuracy of 99.9% but we have many False Positives and False Negatives which makes a false prediction and loss to Bank and customer.

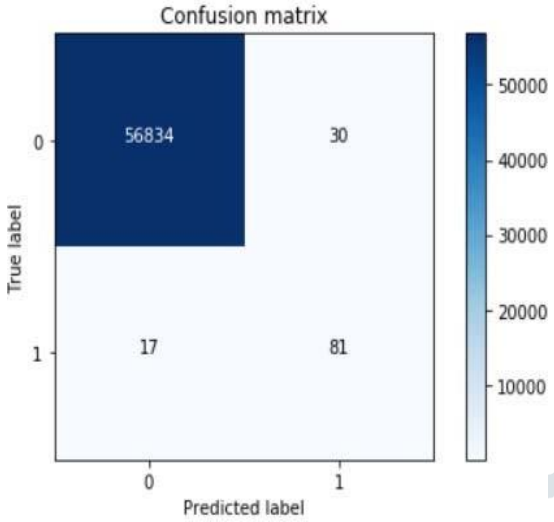

Fig 1.7 -Describes Here Decision tree classifier before using data mining techniques and applied

The above fig shows us high false results. So, As discussed, we why we need to clean data and apply datamining techniques like outlier mining and scaling data.

We can be able to get better results through this we also applied algorithms with some features results in better outcomes.[9]

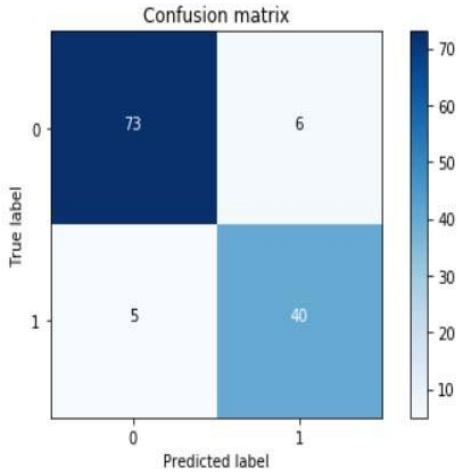

Fig 1.8 DECISION TREE CLASSIFIER

Based on fig 1.8 the algorithm works fine but False positive and False Negatives are bit more we can achieve better accuracy.

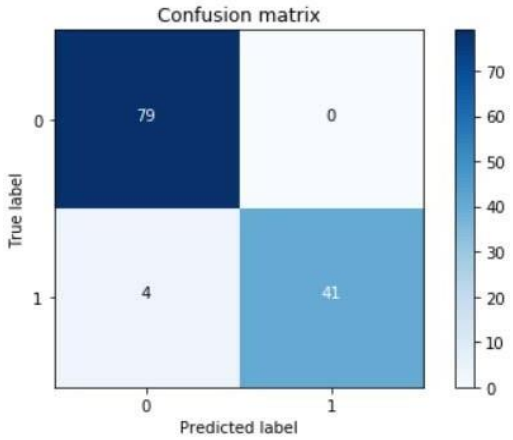

Fig 1.9 RANDOM FOREST CLASSIFIER

Based on fig 1.9 the algorithm works best for False positives and better than fig1.8

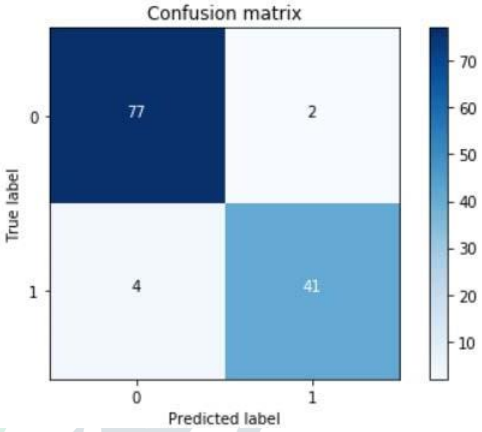

Fig 2.0 LOGIST IC REGRESSION CLASSIFIER

Based on fig 2.0 the algorithm works good but not great than fig 1.9, it also has some false negatives and positives.

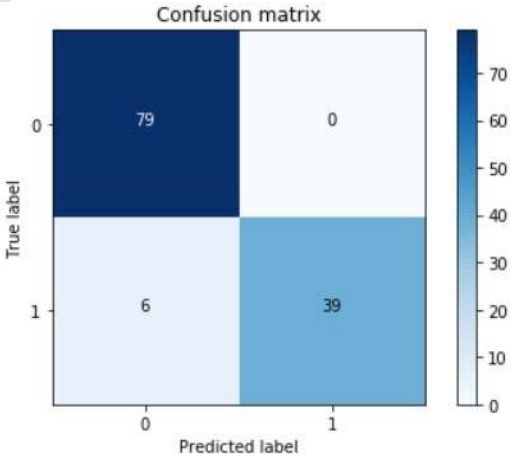

Fig 2.3 NAIVE BAYES CLASSIFIER

Based on fig 2.3 the algorithm works best for false positives but it's not better than fig 2.1.

From fig 2.0 to 2.3 Describes that that how the algorithm works by using confusion matrix and scores which used to get conclusion for deciding which algorithm works best

Table 1- Statistics of Models

	Decision trees classifier	Random forest	Logistic regression	Naive bayes
Accuracy score	91.12%	96.77%	95.16%	95.16
Precision score	86.95%	100%	95.34%	100%
Recall score	88.88%	91.11%	91.11%	86.66%
F1 score	87.91%	95.34%	93.18%	92.85%
Roc-auc score	90.64%	95.55%	94.28%	93.33%

V. CONCLUSION

In this paper we studied the algorithms decision tree , Random forest, logistic regression, naive bayes classification [5] machine learning algorithms , results shows that Random forest classifier performs best with having 96.7741% accuracy , 100% precision , 91.1111% recall , 95.3488% f1 scores and 95.5555 ROU-AUC score and still there are 4 Fa lse Negative values and when we use data without Random Under Sampling we will get accuracy of 99.98% due to heavily imbalance and results many false output. After cleaning data and applying algorithms we got random forest as best algorithm still we can see there is much lee d ifference between all four algorithms but based on the decisions for transactions whether fraud or nor by making a particular feature as root and gaining information from all trees predicting the outcome.

All the algorithms performed almost same with less difference, but we can consider that If these algorithms are trained with some more real-world data then the efficiency and prediction will increase [1]

As we could not reach 100% accuracy even after making many data mining techniques, but we are trying to get more accuracy we will work on combin ing different algorithms [2] which can give us better accuracy and with respect to data also if we get more data and that is by bank officially, we can gain more accuracy as learning increases, we will try to decrease False Negatives.

We can be still ab le to get better results if train models with more data and use genetic algorithms which we will test in our future work.[10]

REFERENCES

[1] R. R. Subramanian, R. Ramar, "Design of Offline and Online Writer Inference Technique", International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2S2, Dec. 2019, ISSN: 2278-3075

[2] Subramanian R.R., Seshadri K. (2019) Design and Evaluation of a Hybrid Hierarchical Feature Tree Based Authorship Inference Technique. In: Kolhe M., Trivedi M., Tiwari S., Singh V. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 39. Springer, Singapore

[3] Joshva Devadas T., Raja Subramanian R. (2020) Paradigms for Intelligent IOT Architecture. In: Peng SL., Pal S., Huang L. (eds) Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm. Intelligent Systems Reference Library, vol 174. Springer, Cham

[4] R. R. Subramanian, B. R. Babu, K. Mamta and K. Manogna, "Design and Evaluation of a Hybrid Feature Descriptor based Handwritten Character Inference Technique," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.

[5] Andrew. Y. Ng, Michael. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", Advances in neural information processing systems, vol. 2, pp. 841-848,2002

[6] John Richard D. Kho, Larry A. Vea " Credit Card Fraud Detection Based on Transaction Behaviour" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[7] Yashvi Jain, Namrata Tiwari, ShripriyaDubey, Sarika Jain, " A Comparative Analysis of Various Credit Card Fraud Detection Techniques, Blue Eyes Intelligence Engineering and Sciences Publications 2019"

[8] Learning Robert A. Sowah, Moses A. Agebure, Godfrey A. Mills, Koudjo M. Kaumudi, " New Cluster Undersampling Technique for Class Imbalance "of 2016 IJMLC

[9] Baraneetharan, E. "Role of Machine Learning Algorithms Intrusion Detection in WSNs: A Survey." Journal of Information Technology 2, no. 03 (2020): 161-173

[10] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 145-152.

[11] Mohamed Jaward Bah, Mohamed Hammad " Progress in Outlier Detection Techniques: A Survey" Hongzhi Wang, of the 2019 IEEE