MACHINE LEARNING ALGORITHAMS TO DETECT THE CREDIT CARD FRAUD BY DOING CAMPARISIONS.

Mansoor Syed
Faculty of Engineering, Environment and Computing
Coventry University
Coventry, England
Syedm35@uni.coventry.ac.uk

***Abstract*—Now-a-days it is common issue around the world for financial institutions to facing credit card defaults. companies are trying to reduce those types of fraud by doing different methods, but comparing all, Machine learning algorithms plays major role to eradicate to this type of problems. The current paper deals with supervised Machine Learning classification algorithms like Decision tree, Random Forest, logistic regression, K-NN and support vector machine learning algorithms to know the accuracy, sensitivity, and precision by testing and training the modules. By comparing all algorithms this paper shows which module is the best for this particular problem. In addition, Python language is used to compare and visualize.**

***Keywords: Python, Decision tree, SVM, Random Forest, K-NN, logistic regression, Machine learning algorithms*.**

I. INTRODUCTION

Early in the 1980s, the usage of credit, debit and prepaid cards are increased enormously around the world [1]. According to Nilson latest report from 2019, In 2018 more than $40.582 trillion were generated worldwide by these payment systems. Cards become basic need to the people to fulfil their daily routines. Recent Nilson report says payment cards used for fuelling station, retail stores, airlines, medical practices and other personal needs combined for $902.60 billion in total volume in 2018.The global wide financial brands are American Express, Dinner club, Master Card, JCB, Visa, Union pay combined card volumes accounted for $33.731 trillion that means nearly 83% total worldwide in 2018. This becomes golden chance to fraudsters to create unwanted data and execute their plans. Credit card frauds are increasing day by day due to easily accessible data around the world in different forms like social media or spam mails sending like a banker and theft personal information. In UK alone in the year 2018 nearly £844.8 million financial loss occurs due to frauds (WIKIPEDIA). worldwide fraud losses increased to $27.85 billion in the year 2018 and it anticipate extending $35.67 billion by 2025 as shown in Fig. 1.[2]

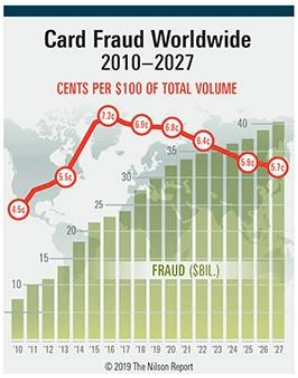

Fig. 1. Global Card Frauds in 2019

Even reputed most powerful credit card companies also facing enormous loss every year. Gross fraud for the international brand card accounted to $24.86 billion in 2018 approximately 89% of gross losses worldwide for all cards. Billions of pounds fraud is happening every annual due to fraudsters [2]. The implementing of efficient algorithm is a primary factor to reduce these types of frauds most of the algorithm's relay on machine learning technics and helpful for fraud investigators.

A. Types Of Frauds

They are many kinds of credit card frauds, and they change according to new technologies enable novel cyber crims that is nearly impossible to list them all in one place [16] some of the important types of frauds shown in Fig.2. [3]

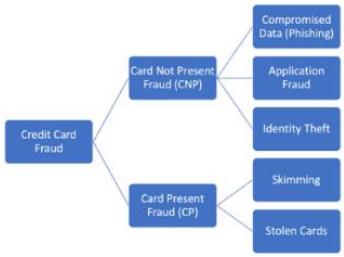

Fig.2.Main Types of Fraud Transactions.

Most important and common frauds are CARD-NOT-PRESENT(CNP) and CARD-PRESNT(CP) frauds.

- CARD-NOT-PRESENT FRAUD: Card-not-present fraud is a type of EMV (Europay, Mastercard, Visa) fraud. It is a type of fraud where card holders don't make transaction in their presence. This type of frauds is difficult to predict because it's impossible for financial institutions to understand whether that card belongs to original customer or not and not able to examine holograms on the cards while in fraud transactions. This process occurs through the technique called PHISING, by using this method fraudsters will trap the customer with spam mails like pretending like they are bankers and collects confidential information. Recent article from FICO, in that article they mention that CNP fraud contain 67% of all frauds in the US. In addition, the data shown from 2014 to 2017 there were gradually increase of CNP fraud from 50% to approximately 70%. This shows that there can be chance of increase in future. problem is growing throughout the globe to stop this type of problems FICO Company uses New Machine learning algorithms and improves 30%[4].

Fig.3 shows using New Machine learning Models Ecommerce fraud was detected more efficiently comparing with old technology [5].

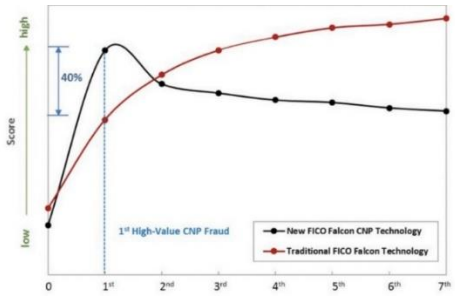

Fig.3. CNP Fraud Detection using New Machine Learning Models.

- CARD PRESENT FRAUD: This type of frauds is very popular today. It's often takes in the form of SKIMMING, when fraudsters place chips in swiping machine to scan and record the personal information as shown Fig.4.

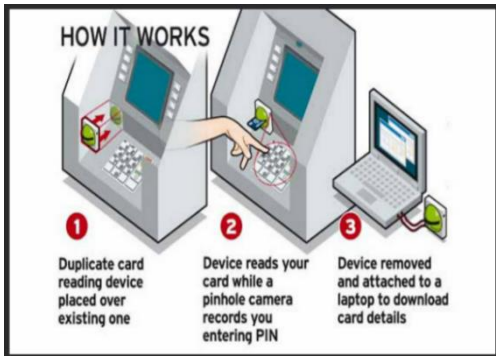

Fig.4. CP Fraud Chip Injecting. [google images]

B. Machine Learning

Machine Learning is a techniques where human intervention is minimal to analyse, identify data and easily make decision through data. machine learning is a branch of Artificial Intelligence, it can predict real world problems by using different algorithms through programming. This feature is useful to detect fraudsters in credit card fraud. By using machine learning model in banking domain, it can reduce risky transactions.

Machine learning is classified into three main categories, but we are using here Supervised Machine learning for this problem.

C. Supervised Mancine Learning

In Supervised learning, machine uses well labelled data to train that means it has corrected answers. It is also known as predictive learning because it generates output based upon past data. in supervised learning the required input will be given to train the data. the independent variable must be past information of the particular problem then only it can predict output and make process.

II. DATASET USED

A dataset which was used in this paper was original data from Kaggle website which was publicly available imbalanced dataset. If the data was imbalance, inconsistency occurs in the output. In this problem data contain two days transactions made by credit card in September 2013 by European card holders. In that set 492 frauds out of 284807 transactions which is extremely unequal percentage. Data contain only numerical inputs and the features used in this paper were PCA transferred from V1 TO V28. Time and Amount is the only attribute which was not transferred with PCA. Class is the important feature which shows 1 if the transaction is fraud and 0 if it is not fraud [15].

A. Dataset Train and Test Procedure

The dataset Train and Test procedure is the main technique for classification algorithms to evaluate performance of each model. This method not suitable for some problems where small amount of data uses because data become imbalance due to splitting technique. Train and Test procedure made by splitting of data into different forms. In this method this technique is performed in python by implementing dependencies from sklearn model selection.

Train and test procedure involve diving a dataset into two subsets [7]. Train dataset and Test dataset, Train data set follows the fit procedure and Test is used to evaluate the fit procedure.

B. Dataset Splitting Procedure

Splitting dataset has some certain procedures but most of the time 70 and 30 makes better results. So, in this project used 70% for training and 30% for testing.

Below some common split percentage include [7].

- Training data = 80%, Testing data =20%
- Training data = 67%, Testing data = 30%
- Training data =50%, Testing data =50%

III. PARAMETERS AND ITS COMPARISION

In order to evaluate the performance of various models, this paper uses various parameters like Accuracy, Sensitivity, Precision, and Time. By testing and evaluating of each algorithm models and comparing parameters results with another models gives the result of which is the best and fastest one to this problem.

A. Accuracy

Measures the proportion of True Positive and True Negative in all evaluated cases. Accuracy is most important deciding factor for the algorithm to know the accurate result of performances but for imbalanced data it cannot act as primary parameter.

Mathematically, accuracy calculated as follows:

$$\text{Accuracy} = TP+TN/TP+TN+FP+FN$$

B. Sensitivity (Recall)

Measures the proportion of True Positive that are correctly identified. In this parameter some portion is positive, but it indirectly calculated as negative. Sensitivity is also known as recall. This parameter plays primary role in this project.

Mathematically, sensitivity calculated as follows:

$$\text{Sensitivity} = TP /TP+FN$$

C. Precision

Measures the proportion of positive predictions that are truly positive.

Mathematically, Precision calculated as follows:

$$\text{Precision} = TP/TP+FP$$

D. Time

Time was one of the primary parameter for this problem to evaluate the performance time of each algorithms and to know time taken by this methods for training and testing the data. comparison made with time for each classification model to understand which was fastest and best for this project.

IV. MACHINE LEARNING MODELS WITH CLASSIFICATION TECHNIQUES

A. Decision Tree

Decision Tree was most popular algorithm in machine learning. A single decision tree can manage the total credit card fraud datasets of a financial institutions [13]. This algorithm is very useful for machine learning because it breakdown huge data into simple parts. It can perform complex task in very quick manner, especially when large amount of data is used to mine. Decision Tree make their choice via a set of binary splits (yes or no answers) from root node through branches passing several internal nodes until it reaches to leaf node where the prediction made [8]. It was extremely useful for big data problems because of the inner working, it can handle large datasets and increase volume of the data without halting their prediction speed or losing their high accuracy. But with huge dataset size the depth of decision tree grows, it makes many if else statements which increases complexity which takes more time to train. Decision Tree look like tree like structure as shown Fig.5[8].

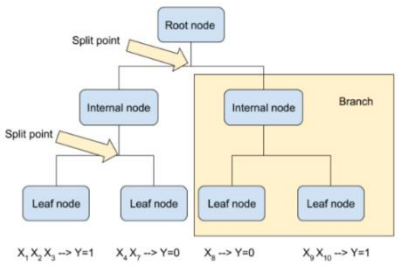

Fig.5. Decision Tree model

B. K-Nearest Neighbour (K-NN)

K-Nearest Neighbour is one of the fastest training model in machine learning. It is also known as lazy

learner model because through learning set it did not respond immediately. Alternately it stores the data and in the time of classification, it performs techniques on dataset. In the time training it stores the data, when new data arrives it understand it places this new data near where similarity matches. an appropriate distance matric is also a requires like metric parameter “Minkowski” also used sometimes in K-NN which decides distance between the points. A suitable value of k should be chosen, however, there is no mathematical formula to determine a KNN algorithm to classify the given dataset, k value is chosen by experiment with different values. But the most appropriate k value is 5 [15]. Note this algorithm uses the Euclidean distance to describe the similarity between two elements is measured.

The Euclidean distance between two input vectors

$$D = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

C. *Random Forest Algorithm*

Random Forest is based on the concept of ensemble learning. This model uses more decision tress to be averaging the data by taking inputs from n-number of training data. In this model prediction made on the basis of polling.

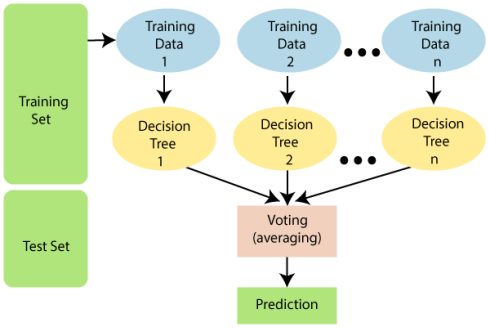

Fig.6. Random Forest Algorithm

D. *Logistic Regression*

Logistic Regression is one of the most popular classification algorithm. In this model, output can be generated based on the input variable. The results always discreate or categorical values like yes or no, True or False, 0 or 1. But it did not ex the exact vale 0 or 1 but it must be between them, so it forms curve like S shaped called sigmoid function.

Logistic Regression classified in three types of Binomials (two possible dependents), Multinomial (3 or more possible dependents) and Ordinal (3 or more possible dependents). [9]

Equation of Logistic Regression

$$\text{Log}(y/1-y) = b0+b1x1+b2x2+\text{----------}+bnxn.$$

E. *Support Vector Machine*

This supervised machine learning algorithm aim is to find an optimal hyperplane to divide the dataset into different classes in a n-dimension space. If the data are not linearly separable, SVM algorithm can pre-process the data and represent the features in a higher-dimensional space in which they can become linearly separable, Kernel function also used in this method to map the data into higher dimensions. Support vectors are crucial data points that helps to optimize the hyperplane. These vectors lie so close to the hyperplane and difficult to classify. the position of the hyperplane mainly depends on the support vectors, if support vectors are removed then it will also change the position of the hyperplane. To classify the given dataset, a linear and gaussian kernel are used, with the linear kernel provides much better results for this data set [11].

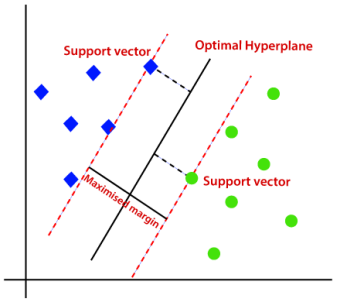

Fig.7. Support Vector Machine

V. APPLICATION TECHNIQUES AND RESULTS

In this project algorithms were Executed with Python language in Jupiter Notebook 6.3.0 in Anaconda Navigator by using Lenovo laptop with Intel core i5, 2.4GHz RAM 8 GB.

Code extracted from different sources of the platform and modified as project requires, Models were trained using Parameters that delivers the best results, Parameters were obtained by using grid search. All the models were trained by using 28 features from V1 to V28. each model trained more than five times make sure to get best result below TABLE 1 shows Training time taken by these models.

TABLE 1 TRAINING TIME

Models	**Training Time**
Decision Tree	28.79s
K-NN	0.035s
Random Forest	177.59s
Logistic Regression	1.55s
SVM	5334.07s

Accuracy, F1 Score, Recall, Precision and Confusion Matrix were obtained for all models and presented each model separately. In addition, ROC and AUC Curve also obtained, respectively.

Confusion Matrix plays major in this project we use to understand mainly Precision and recall in all algorithms to make predictions sample diagram below simple Fig.8.

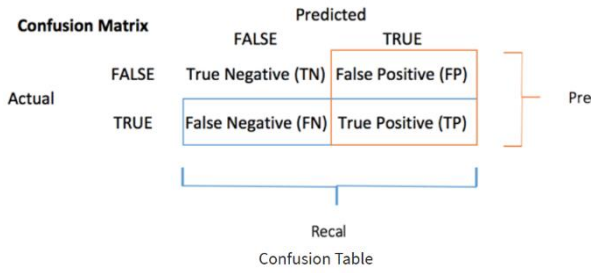

Fig.8 Confusion Matrix [12].

A. *Decision Tree*

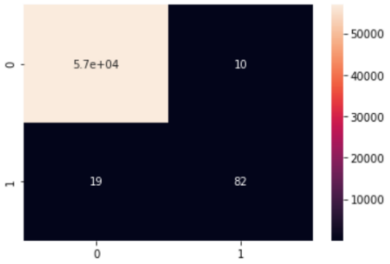

Fig.9 Confusion Matrix of Decision Tree.

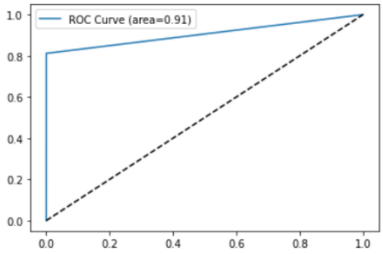

Fig.10 ROC_AUC Curev of Decision Tree.

TABLE 2 DECISION TREE RESULTS

Parameters	Results	Confusion Matrix	RESULTS
Accuracy	99.4%	TP	56850
Precision	88%	TN	78
Recall	77%	FP	11
F1 Score	82%	FN	23

B. *K-Nearest Neighbour(K-NN)*

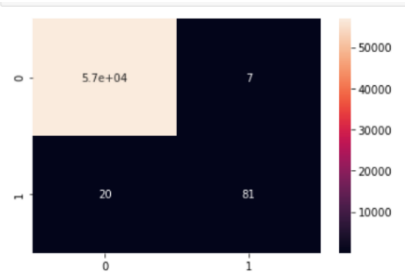

Fig.11 Confusion Matrix of K-NN.

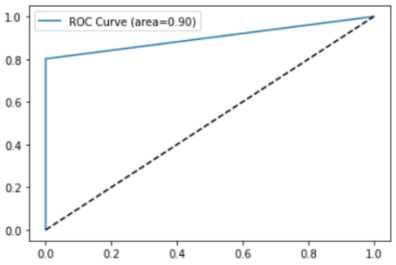

Fig.12 ROC_AUC Curve of K-NN.

TABLE K-NN RESULTS

Parameters	Results	Confusion Matrix	Results
Accuracy	99.5%	TP	56854
Precision	92%	TN	81
Recall	80%	FP	7
F1 Score	85%	FN	20

C. *Random Forest*

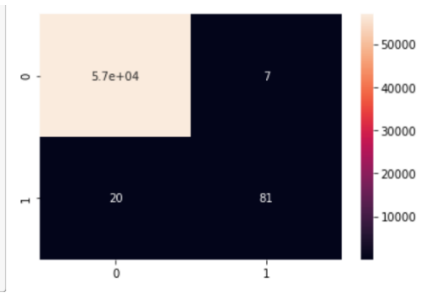

Fig.13 Confusion Matrix of Random Forest

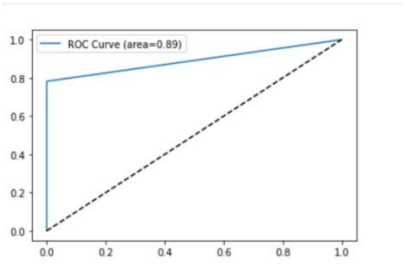

Fig.14 ROC_AUC Curve of Random Forest.

TABLE 4 RANDOM FOREST

Parameters	Results	Confusion matrix	Result
Accuracy	99.4%	TN	56855
Precision	93%	TP	78
Recall	77%	FN	23
F1 Score	84%	FP	6

D. Logistic Regression

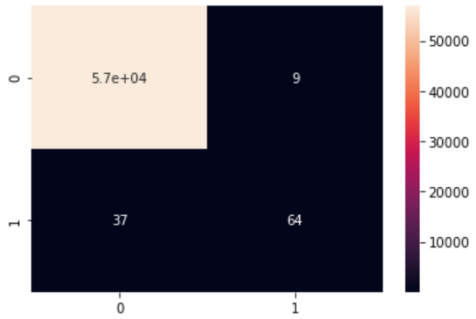

Fig.15 Confusion Matrix of Logistic Regression

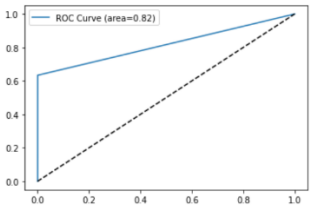

Fig.16 ROC_AUC Curve of Logistic Regression.

TABLE 5 LOGISTIC REGRESSION RESULTS

Parameters	Results	Confusion Matrix	Results
Accuracy	99.1%	TN	56852
Precision	88%	TP	64
Recall	63%	FN	37
F1 Score	73%	FP	9

E. Support Vector Machine

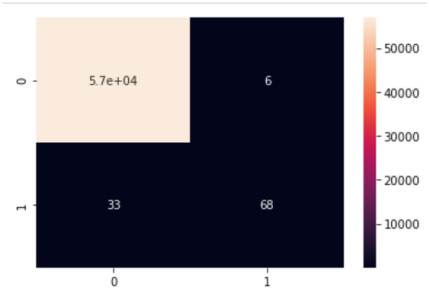

Fig.17 Confusion Matrix of SVM

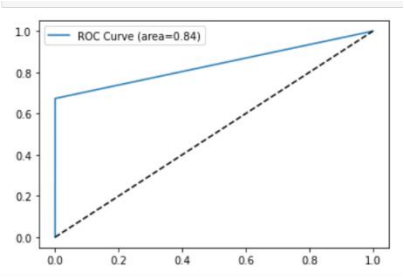

Fig.18 ROC_AUC Curve of SVM.

TABLE 6 SVM RESULTS

Parameters	Results	Confusion Matrix	Results
Accuracy	99.3%	TN	56855
Precision	92%	TP	68
Recall	67%	FN	33
F1 Score	77%	FP	6

VI. OBSERVATIONS

By observing all Classification methods, the accuracy score is very high showing more than 99% due to imbalanced dataset, but Sensitivity and Precision rates are gradual changes so it can be useful to understand which is the best for this problem, if you see above algorithms results Decision Tree and Logistic Regression Precision result are same at 88% and Decision Tree Sensitivity is 14% higher than Logistic Regression was 77% and 63% respectively. Random Forest has high percentage in precision comparing all at 93% but its recall value is 77% only. K-NN has lead performance in both sensitivity and precision, 92% and 80% respectively and it is the fastest training model by 0.035s s comparing total algorithms.

We also displayed Confusion Matrix by using Python Coding and obtain TN, TP, FN, and FP. We also drawn ROC_AUC_CURVE by using default threshold values (0.5). At last training time was calculated.

VII. CONCLUSION

In this project we used five types of Machine Learning Classification Algorithms k-NN, Decision Tree, SVM, Logistic Regression and Random

Forest to know the fraud detection in credit card transaction, this is very common issue around the world, government and financial institutions are trying to reduce this problem with different efforts, but machine learning methods has prominent role to overcome these issues. So, we used best machine learning methods, comparing all we understand k-NN get more accurate result of Sensitivity and Precision value than four algorithms, and also Training Time is very less than all the models, so we choose K-NN is the best over four algorithms because of fast Execute time and also best results. To ensure that minimum time to be taken for prediction, therefore, K-NN is the best model. Here we are not considering accuracy as primary parameter due to imbalance in data.

VIII. REFERENCES

[1] B. Buonaguidi, "Credit card fraud: What you need to know," *BBC Work life,* 2017.

[2] "Card Fraud Losses Reach $27.85 Billion," *Nilson Report,* 2019 November.

[3] D. Bhoopesh, "Is your credit card is safe ? know more about your credit card fraud and safety," *Forensic Science for Healthcare Professionals,* 2019.

[4] S. Zoldi, "Machine Learning Improves CNP Fraud Detection Rates by 30%.," *FICO,* 2O18 APRIL 3.

[5] K. Students, "Can machine learning provide a solution to CNP frauds?," *A project of Digital Initaiztive by Hardvard business school.*

[6] "Machine Learning," *IBM CLOUD.*

[7] J. Brownlee, "Train-Test Split for Evaluating Machine Learning Algorithms," *Machine Learning Mastry,* 2020 august.

[8] "The Ultimate Guide to Decision Trees for Machine Learning," *Keboola,* 2020 september 10.

[9] "Random Forest Algorithm," *JAVA T POINT.*

[10] "Logistic Regression in Machine Learning," *JAVA T POINT.*

[11] "Support Vector Machine Algorithm," *JAVA T POINT.*

[12] "Confusion Matrix in Machine Learning with EXAMPLE," *GURU99.*

[13] R. Kazmi, "Credit Card Fraud Detection: Machine Learning at its Best," *Koombea,* 2021 march 10.

[14] "Credit Card Fraud Detection," *kaggle,* 2019.

[15] "K-Nearest Neighbor(KNN) Algorithm for Machine Learning," *JAVA T POINT.*

[16] "Most Common Types of Fruads," Australian Cyber Security Services.

[17] A. Das, "How To Choose The Best Machine Learning Algorithm For A Particular Problem?," Analytics India Magzine, 2020 October 18.

APPENDIX-PYTHON PROGRAMING LINK

https://drive.google.com/file/d/1hZtLwitgujGgS0XLd5A0msL3u0-d8ZaD/view?usp=sharing