2nd National Conference on Engineering Applications of Emerging Technology
in association with International Journal of Scientific Research in Science,
Engineering and Technology | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
[https://ijsrset.com | doi : https://doi.org/10.32628/IJSRSET]

Analysis of Fraud Detection using Computational Algorithms

Darshan S Aladakatti, Gagana P, Dr. Ashwini Kodipalli
Department of CSE, Global Academy of Technology Bangalore, Karnataka, India

ABSTRACT

Electronic fraud is a type of crime that involves the use of various forms of financial instruments such as credit cards, insurance claims, and cell phones. Due to the complexity of these transactions, specialized techniques are needed to identify fraud using data mining, machine learning, and knowledge discovery. These tools are able to provide businesses and governments with effective solutions to their problems. One of the main reasons why data analytics is used to identify fraud is due to the weaknesses in internal control systems. This is because many companies have serious issues with their systems. For instance, law enforcers typically rely on circumstantial evidence to detect potential fraud. Due to the lack of effective tools and techniques to detect fraud, many cases of this type of crime remain undetected. To effectively address these issues, organizations and businesses can use data analytics to analyse and monitor their control systems. These tools can help them detect and prevent unauthorized transactions.

Keywords-Fraud, Machine Learning, Analysis, LR, KNN, SVM, Naïve Baye's, Decision Tress, Random Forest

I. INTRODUCTION

In finance banks analyse their past data to build models to use in credit application, fraud detection and the stock market. The World wide web is huge it is constantly growing, and searching for relevant information cannot be done manually. However machine learning is not just a database; it is also a part of artificial intelligence. Machine learning system is exposed to changing environment in which it will learn and adapt to changes and provide solutions for all possible situations. [1]Machine Learning is programming the computers to optimize the performance criterion using example data or past experience we have a model up to some parameters, and learning is the execution of a computer program to optimize parameter of the model using the training data or past experience. The model may be predictive to make predictions in the future or descriptive to gain the knowledge from data, or both. Banks could benefit from a machine learning-based fraud detection solution in that they would be able to instrument it across more than one channel of data to be analysed.

We, as engineers are determined to predict fraud, using ML algorithms and data analytics to prevent dodging into fraud. We use different ML algorithms and evaluate their performance and analyse the dataset. These algorithms can also be applied to different datasets can also be implemented at a further stage.

The manuscript is organized as follows:

- Literature survey

- Description of the dataset.
- Implementation and Result discussion of ML algorithms.
- Conclusion-future work

II. LITERACY SURVEY

Recently, AL, ML and DL has been used extensively in finances. Basically, data mining classification comprises problems such as Fraud detection that is used to figure out online transactions as fraudulent or legitimate. Some Additional techniques and factor methods apart from data mining which are involved in fraud detection are Web- services based collaborative schemes in which the private bodies like banks can share the information about the fraud patterns and frequencies for the enhancement of the fraud detection capability and to reduce the financial loss. The basic procedure in developing any Machine Learning model [2] is given below in figure 1.The comparison among various models and their resultant analysis, like XGBOOST, Random Forest, Decision Trees etc. These are the most widely used techniques so far in detection of frauds. There has also been a study of new techniques like Adaboost and Majority Voting approaches that add or enhance the ML algorithm performance.[3,4]Fraud detection in simple words is a simple binary classification problem in which any particular transaction or exchange will be either classified as fraud or legit only. In this study, a few standard classification techniques like Naive Bayes, K-Nearest Neighbor and Logistic Regression methods, Random Forest Classifiers, Decision Trees. For effective usage of these algorithms, different stages are included such as gathering the data, cleaning data, researching and visualization of data and training the classifier algorithms and finally evaluating the result.[8]

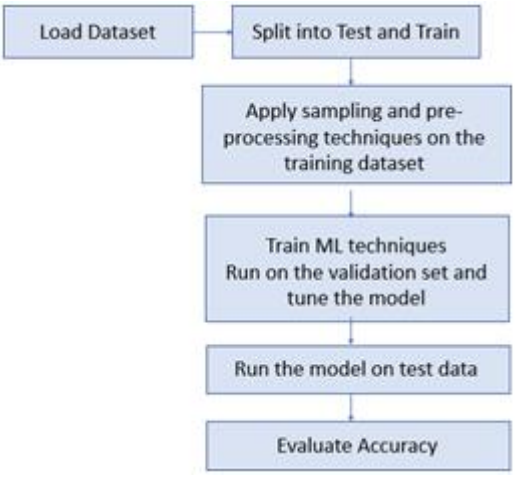


Fig 1: The flow process diagram for developing a machine learning model

III. DESCRIPTION OF THE DATASET

Lately as the result of the world getting digitalised there is an exponential utility of online transaction, which has indeed led to higher number of fraudulent activities. Approximately 49% of digital transactions are victimized to fraud. In this paper, we used Kaggle dataset that predicts whether the transaction is likely to be fraud

This works on the basis of the inputs i.e.,

step: represents a unit of time where 1 step equals 1 hour type: type of online transaction

amount: the amount of the transaction nameOrig: customer starting the transaction oldbalanceOrg: balance before the transaction newbalanceOrig: balance after the transaction nameDest: recipient of the transaction

oldbalanceDest: initial balance of recipient before the transaction

newbalanceDest: the new balance of recipient after the transaction isFraud: fraud transaction[10]

IV. ML ALGORITHMS US ED

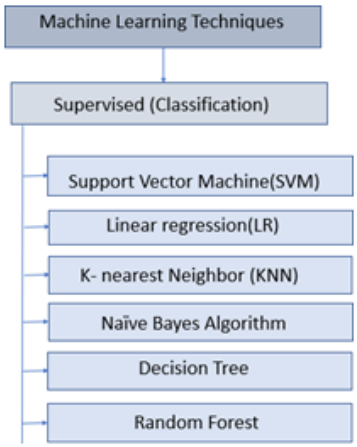
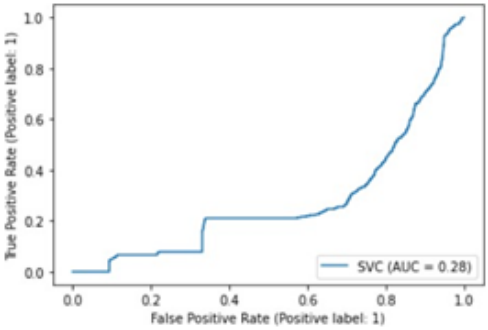


Fig 2- ML classification algorithms

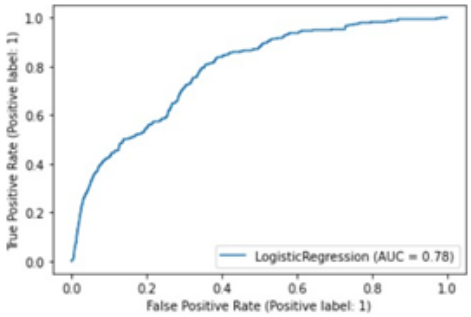
A. SVM ALGORITHM

One of the most widely used support vector machine (SVM) algorithms in machine learning is the K-Nearest Neighbor. It can perform both classification and regression problems. In order to solve these problems, the algorithm takes into account the various requirements of the system and then divides the data points into classes.



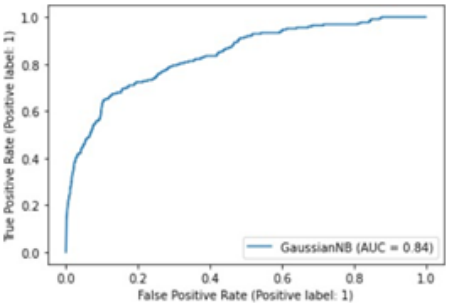
B. LR ALGORITHMS

Logistic Regression (LR) is an algorithm that is used to accurately predict the value of a data member. It is also a classification algorithm. The datasets used here are labelled and hence the algorithm is a supervised algorithm. Since it is a type of classification algorithm, the outcome is only 2 values, either a true (1) or a false (0) / yes (1) or a no (0). [5]



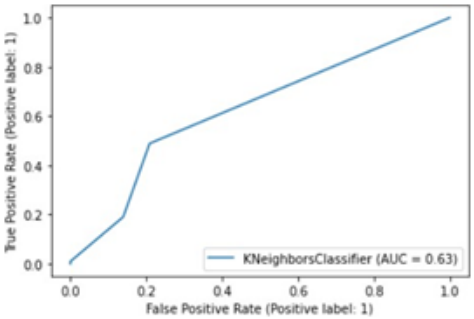
C. NAIVE BAYES

Naive Bayes works on the principle of conditional probability, as given by the bayes theorem. It falls under classification supervised machine learning algorithm.It calculates the conditional probability of one event assuming that the other event has already occurred. This algorithm is mainly used in face recognition, weather prediction etc.



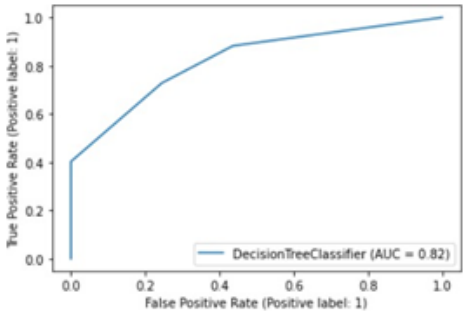
D. KNN ALGORITHM

K-Nearest Neighbor is a machine learning algorithm that can perform both classification and regression problems. It takes into account the distance between the two neighboring regions and then computes the optimal number of K nearest neighbors.



E. DECISION TREE ALGORITHM

A decision tree is a type of algorithm that is used to solve both classification and regression problems. It starts with a root node and then moves to a series of if-else like statements. The child nodes and the leaf nodes are also if- else like statements.



F. RANDOM FOREST ALGORITHM

Random forest algorithm is also used to resolve both classification and regression problems. It's a classifier that takes the average of a succession of decision trees for distinct subsets of a given dataset to enhance the dataset's prediction accuracy. In comparison to previous algorithms, less trainingtime is required. It bodes well with large datasets and can accurately generate outputs[7]

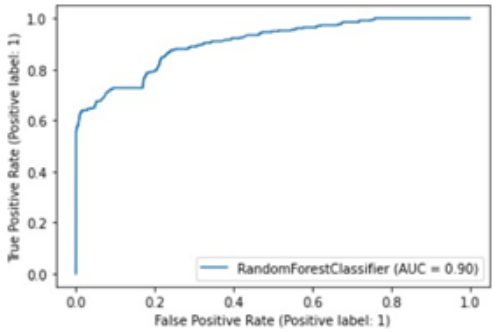


Fig. 1. Comparative study using ROC Graph

Figure above is the comparative study that shows us the AUC Graph for each of the algorithms used in Fraud detection. On X-axis, we have the false positive rate and on xis, we have the true positive rate. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. A Receiver Operating Characteristic curve outlines a classification model's achievement across all levels. The true positive rate and the false positive rate is mapped on this curve[8]

G. TABLES

Parameters	SVM	LR	Naïve Bayes	KNN	Decision Tree	Random Forest
Accuracy	55.41%	95.73%	98.61%	97.10%	99.94%	99.94%
Precision Score	0.000499	0.007226	0.022032	0.001705	0.014709	0.954802
Recall Score	0.210210	0.288288	0.276785	0.045248	0.615384	0.50750
F1 score	0.000997	0.014100	0.040816	0.003287	0.028731	0.662745
Log-loss	15.39865	1.474041	0.479928	0.998561	0.021904	0.018884

Table: Accuracy, Precision Score, Recall Score and F1 Score when different algorithms used

Accuracy: A standard metric that identifies which model isbetter at detecting patterns between the variables andcorrelations between the variables in a dataset based on the input, or training data. It is also known as the proportion of accurately anticipated observations to all observations.

Accuracy = TP+TN/TP+FP+FN+TN;

TP = True Positives, TN = True Negatives, FP = False Positives, FN = FalseNegatives.

Precision Score: The ratio of accurately predicted positive observations to the total number of expected positive observations are the Precision Score. A low false positive rate is associated with high accuracy.

Precision = TP/(TP+FP)

Recall Score: It is the proportion of accurately anticipated positive observations to the total number of observations in the class. The Recall value must be greater than 0.5 at all times.

Recall Score = TP/(TP+FN)

FI Score: The F1 score is a significantly better assessment of the model since it is the harmonic mean of Precision and Recall. A solid F1 score indicates that you have a low number of false positives and negatives.

F1 Score = 2*((precision*recall)/(precision+recall)). Confusion Matrix: The performance of a classification algorithm is condensed using a confusion matrix. If there is an imbalanced number of observations in each class or if the dataset consists of multiple classes,the accuracy alone might be not be enough. We can obtain a clearer picture of what our classification model gets correct and what sorts of mistakes are being created by utilizing a confusion matrix.

TP: True Positives means the predicted positive value and actual positive value is same. Example: Tapturns on when bucketis empty.

TN: True Negatives means the predicted negative value and actual negative value is same. Example: Tap is turned off when bucket is notempty.

FP: False Positive refers to when the Original value of a classis negative but the anticipated value of a class is positive. Example: Tapturns on when bucket is not empty.

FN: False Negative refers to when the original value of a classis positive but the anticipated value of a class is negative. Exa mple: Tap is turned off when bucket is empty

H. Confusion matrices for all algorithms used (3x3 matrix)

SVM ALGORITHM

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=174258	FP=139982
Original YES	FN=263	TP=70

LR ALGORITHM

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=301052	FP=13188
Original YES	FN=237	TP=96

NAÏVE BAYES ALGORITHM

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=206739	FP=2725
Original YES	FN=162	TP=62

KNN ALGORITHM

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=209493	FP=1
Original YES	FN=221	TP=0

DECISION TREE ALGORITHM

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=209493	FP=1
Original YES	FN=132	TP=89

RANDOM FOREST

n=1048576	Anticipated: NO	Anticipated: YES
Original NO	TN=314232	FP=8
Original YES	FN=164	TP=169

V. CONCLUSION

In this paper, we have tried to analyze the accuracy of various fraud detection algorithms. We have also considered the various confusion matrices used in these algorithms.Random Forest among other algorithms is giving better results when compared to the other algorithms with an accuracy of 99.94%.In order to improve the accuracy of fraud detection algorithms, more research is needed on analyzing the various data collected by machine learning systems.Hence, some pre-processing and sampling algorithms must be added to the raw dataset to classify it before being subjected to any of the SVM, LR, Naïve Bayes, KNN, Decision Tree, Random Forest, techniques used to analyze it.The main reason why a model might fail is due to the fact that it doesn't protect the sensitive data that it collects while it's in the process. Also, since the number of frauds is always less than the total data that it considers, the variation in the model's subset can cause it to fail in a particular situation.To prevent this, we would greatly improve the accuracy of our model by implementing it through a sophisticated front end, which can be accessed through Flask. In addition to this, we would also explore other methods of machine learning to improve its prediction.

VI. REFERENCES

[1]. Introduction to machine learning, third edition by Ethem AlpaydinProfessor Department of computer engineeringBogazici university, Istanbul pg. 3

[2]. Vinod Jain, Mayank Agrawal, Anuj Kumar " Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection 2020 at 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (2022)

[3]. A. Krishnaiah, P.B. Divakar Chari Automatic Music Classification using Multi-class Support Vector Machine based on Hybrid Spectral Features (2022)Google Scholar

[4]. Gurum urthy Krishnamurthy Arun*, Kaliyappan VenkatachalapathyIntelligent Feature Selection with Social Spider Optimization Based Artificial Neural Network Model for Credit card Fraud detection| Arun & Venkatachalapathy, 11, IIOABJ (2020), pp. 85-91

[5]. Uddin, S., Khan, A., Hossain, M. et al. 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281

[6]. P. Soucy and G. W. Mineau, 2001, "A simple KNN algorithm for text categorization," Proceedings 2001 IEEE International Conference on Data Mining, pp. 647- 648, doi: 10.1109/ICDM.2001.989592

[7]. Kausthubh Priyan, Pavan Kumar KN, Manish HR, Snigdha Sen, " Stroke Prediction and Analysis usin g Machine Learning", ICISSI, 2022, Taylor and Francis Journal

[8]. M J Madhurya , H L Gururaj , B C Soundarya , K P Vidyashree, A B Rajen dra , Exploratory Analysis of Credit Card Fraud Detection using Machine Learning Techniques, Glo bal Transitions Proceedings (2022), doi: https://doi.org/10.1016/j.gltp.2022.04.006

[9]. Blogs: how to create a Fraud detection prediction model: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve- machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes.

[10]. Kaggle dataset- https://www.kaggle.com/datasets/rupakroy/online-payments-fraud- detection-dataset