International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST207671

Credit Card Fraud Detection Framework – A Machine Learning Perspective

Jasmin Parmar*[1], Dr. Achyut C. Patel[2], Dr. Mayur Savsani[3]
*[1]Saurashtra University, Rajkot, Gujarat, India
[2]SMT. M. T. Dhamsania College of Commerce, Rajkot, Gujarat, India
[3]Symbiosis Statistical Institute, Symbiosis International (Deemed University), Pune, Maharashtra, India

ABSTRACT

Article Info
Volume 7, Issue 6
Page Number: 431-435
Publication Issue :
November-December-2020

The short improvement withinside the E-Commerce enterprise has caused a dramatic enlargement withinside the usage of credit score playing cards for on-line buys and thusly they had been flooded with the fraud diagnosed with it. As of late, for banks has gotten extraordinarily tough for figuring out the fraud with inside the credit card framework. Machine getting to know assumes an essential component in distinguishing credit card fraud withinside the transactions. For foreseeing those transactions banks make use of specific system getting to know methodologies, beyond data has been accrued and new highlights are being applied for enhancing the prescient force. The exhibition of possible threats identification in credit card instances is highly prompted through the analysing technique at the informational collection, the dedication of factors, and discovery strategies applied. This paper explores the presentation of K-Nearest Neighbor, Decision Trees, Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost for credit card fraud detection. Dataset of credit card transactions is accrued from Kaggle and it includes a sum of 2,84,808 credit card transactions of an EU financial institution dataset. It depicts doubtful transactions as fraud & labels it "high-quality class" and actual ones as the "poor class". The dataset is relatively imbalanced, it has approximately 0.172% of fraud cases and the relaxations are actual transactions. These methods are implemented for the dataset and work is carried out in Python. The presentation of the methods is classed relying on the accuracy and F1 rating and confusion matrix. Results display that every set of rules may be used for credit card fraud detection with excessive precision. The proposed version may be helpful for the invention of numerous anomalies.

Article History
Accepted : 15 Dec 2020
Published : 30 Dec 2020

Keywords: Fraud detection, K-Nearest Neighbor (KNN), Decision Trees, Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost

I. INTRODUCTION

With the growth and acceleration of e-commerce, plastic cards have been used gigantically for online shopping, resulting in a high level of cheats identified with credit cards. Today is the digital age and the need to distinguish fraudulent transactions of credit cards are essential.

Fraud identification includes checking and investigating the behaviour of customers to assess the differentiation or circumvention of disruptive behaviour. To successfully differentiate credit card fraud detection, we have to go through basics of how a credit card is used, what are the types & usage area of the credit card etc.

Algorithms may or may not separate fraudulent transactions. If you discover blackmail, they must pass record and information about false transactions. They dissect the data set and characterize all transactions.

II. LITERATURE REVIEW

You Dai, et. al [2] truly describe Random forest set of rules relevant to detect frauds. Random forest area has sorted, as an instance, random tree primarily based random forest and CART based random forest area. They depict in the element and their accuracy of more than 91% and 96% separately. The paper also concludes the second type is far better than the number one sort.

Suman Arora [3] stated that several supervised systems which study algorithms generally trail on 70+% training and 30+% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree, naïve Bayes and KNN algorithms reflect on consideration on each other as an instance ˜94.60%, ˜95.30%, ˜94.60%, ˜93.20%, ˜90.90%, ˜90.50% and 94.30% respectively.

Kosemani Temitayo Hafiz [4], they depict move define of the fraud detection process. as an instance information Acquisition, information pre-processing, information evaluation and techniques or algorithms are in the element. Algorithms are K-nearest neighbour, random tree, AdaBoost and Logistic regression accuracy are ˜96.90%, 94.30%, 57.70% and 98.20% individually.

Fraudulent physical games are inflicting sizeable misfortune, which roused professionals to find out a solution that might understand and stop fraud. A few strategies have simply been proposed and tried. Some of them are quick evaluated underneath.

III. PROPOSED TECHNIQUE

The proposed strategies are applied in this paper, for distinguishing the cheats in credit card framework. The correlation is made for different gadget mastering set of rules, Decision Trees, Logistic Regression, Support Vector Machine, Random Forest, and XGBoost to determine which set of rules offers fits satisfactory and may be adjusted via way of means of credit card shippers for distinguishing The Figure1 indicates the constructing graph for speaking me to the in standard framework structure.

The getting ready steps are mentioned in Table 1 to differentiate the satisfactory set of rules for the given dataset.

TABLE 1
PREPARING STEPS

	Algorithm Steps
Step 1	Importing the required packages into our python environment
Step 2	Importing the data
Step 3	Processing the data to our needs and Exploratory Data Analysis
Step 4	Feature Selection and Data Split

Step 5	Building six types of classification models
Step 6	Evaluating the created classification models using the evaluation metrics

Decision Tree is a computational device for type and prediction. A tree incorporates inner nodes which represent a take a look at on an attribute, every department indicates a final result of every leaf node (terminal node) holds a category label. It recursively parcels a dataset making use of both profundities first grasping technique or breadth grasping technique and forestalls whilst all of the factors were appointed a particular for the parcel rule to be talented it must isolate the statistics into bunches in which a solitary elegance prevails in every gathering. As such, the satisfactory parcel may be the only wherein the subsets do not cowl for instance they may be unmistakably disjoint to a maximum excessive sum.

Support Vector Machine is a supervised mastering set of rules wherein given a dataset it isolates them into diverse lessons making use of a hyperplane. The goal of SVM is to find out this hyperplane. There will sever a hyperplane but we're resolved to discover a perfect hyperplane. The focuses nearest to the hyperplane withinside the diverse lessons are referred to as support vectors and the said support vectors are applied to expect the lessons of the latest statistics. To make it more clear, our gadget we feed supervised facts for instance facts with effects without a doubt which are known. It acquires the behaviour of fraud and actual transactions and then it can organize new transactions with admire to which elegance it has an area.

K- Nearest Neighbor (KNN) is possibly the most used algorithm for each type and regression prescient issues. Its performance is based upon three factors: the space measurements, the space rule and the estimation of K. Distance measurements offer the degree to discover closest associates of any coming near statistics factor. Distance rule helps us to represent the novel statistics factor into the arena via way of means of contrasting its spotlights and that of statistics concentrates on the K estimation. A diagram depending on the validation blunders curve is depicted on the graph to perform the estimation of K. It must be applied for almost predictions. We compute the fundamental elegance withinside the vicinity of any new transaction and classify the transaction to have an area with that winning elegance.

Logistic regression is possibly the maximum well-known type set of rules in gadget mastering. The logistic regression version portrays connection among signs that may be constant, binary, and categorical. There are chances that a dependent variable can be binary. Because of positive predictors, we foresee if something will occur. We gauge the chance of getting an area with each elegance for a given association of predictors.

Random forest area is a set of rules that may be applied in each type and regression issues. It incorporates of severing a decision tree. This set of rules offers higher consequences whilst there may be a better wide variety of decision trees withinside the forest area and forestalling version to overfitting. Every choice tree in decision tree offers some consequences. These consequences are blended to get extra particular and solid expectation.

IV. OUTCOME AND ANALYSIS

To figure out the best algorithm is generally appropriate for the issue of distinguishing fraud instances, various measures for algorithm checking has been utilized. Often utilized measurements for deciding the consequences of ML algorithms are Precision and F1 Score. The entirety of the referenced measurements can be determined from a Confusion matrix.

Since here, the test set comprises of 20+% of the dataset, a total summation of samples is 56962. From the all above samples, 101 are fraud transactions, the Decision Tree model accomplished:

Accuracy score: 99.93%
F1 score: 81.05%

TABLE 2
CONFUSION MATRIX FOR DT

Actual	Predicted	
	0	1
0	56849	12
1	24	77

KNN model got the following outcomes (Table 3):

Accuracy score: 99.95%
F1 score: 85.71%

TABLE 3
CONFUSION MATRIX FOR KNN

Actual	Predicted	
	0	1
0	56854	7
1	21	81

LR model got the following outcomes (Table 4):

Accuracy score: 99.91%
F1 score: 73.56%

TABLE 4
CONFUSION MATRIX FOR LR

Actual	Predicted	
	0	1
0	56852	9
1	37	64

SVM model got the following outcomes (Table 5):

Accuracy score: 99.93%
F1 score: 77.71%

TABLE 5
CONFUSION MATRIX FOR SVM

Actual	Predicted	
	0	1
0	56855	6
1	33	68

RF model got the following outcomes (Table 6):

Accuracy score: 99.92%
F1 score: 77.27%

TABLE 6
CONFUSION MATRIX FOR RF

Actual	Predicted	
	0	1
0	56854	7
1	33	68

XGBoost model got the following outcomes (Table 7):

Accuracy score: 99.94%
F1 score: 84.49%

TABLE 7
CONFUSION MATRIX FOR XGBoost

Actual	Predicted	
	0	1
0	56854	7
1	22	79

As indicated by the accuracy score assessment metric, the KNN model uncovers to be the most precise model and the Logistic relapse model to be the most un-precise model. In any case, when we gather together the consequences of each model, it shows 99% precise which is a generally excellent score.

The positioning of the models is practically like the past evaluation metric. On-premise of the F1 score evaluation metric, the KNN model grabs the primary spot again and the Logistic regression model remaining parts to be the most un-exact model.

While contrasting the confusion matrix of the multitude of models, it tends to be seen that the K-Nearest Neighbors model has played out generally excellent employment of classifying the fraud transactions from non-fraud transactions followed by the XGBoost model. So we can infer that the most fitting model which can be utilized for our case is the K-Nearest Neighbors model and the model which can be ignored is the Logistic relapse model.

V. CONCLUSION

Credit card fakes speak to an intense business issue. These frauds can prompt immense misfortunes, both business what's more, individual. Therefore, organizations contribute more and more cash in growing ground-breaking thoughts and the roadmaps that will give us an edge to identify and minimise frauds.

The fundamental objective of the given paper is to think about different ML algorithms for identification of fraudulent transactions. Later, the examination concluded that KNN gives the best outcomes for the given example and gives the exact classification of whether transactions are fraud or not. The set up utilizing various evaluation metrics, for example, precision and F1 Score. Selection of features and dataset balancing have demonstrated to be critical in accomplishing critical outcomes.

The future work should be contributed towards finding out about resampling strategies that will support us with decreasing skewness proportion of the datasets and apply deep learning procedures.

VI. REFERENCES

[1] Naik, Heta, and Prashasti Kanikar. "Credit card fraud detection based on machine learning algorithms." Int J Comput Appl 182.44 (2019): 8-12.

[2] Dai, You, et al. "Online credit card fraud detection: A hybrid framework with big data technologies." 2016 *IEEE Trustcom/BigDataSE/ISPA*. IEEE, 2016.

[3] Arora, Suman, and Dharminder Kumar. "Selection of optimal credit card fraud detection models using a coefficient sum approach." *2017 International Conference on Computing, Communication and Automation (ICCCA).* IEEE, 2017.

[4] Hafiz, Kosemani Temitayo, Shaun Aghili, and Pavol Zavarsky. "The use of predictive analytics technology to detect credit card fraud in Canada." *2016 11th Iberian Conference on Information Systems and Technologies (CISTI).* IEEE, 2016.

Cite this article as :

Jasmin Parmar, Dr. Achyut C. Patel, Dr. Mayur Savsani, "Credit Card Fraud Detection Framework - A Machine Learning Perspective", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 7 Issue 6, pp. 431-435, November-December 2020. Available at doi : https://doi.org/10.32628/IJSRST207671
Journal URL : http://ijsrst.com/IJSRST207671