International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Credit Card Fraudulent Detection using Machine Learning

Kalpesh Vishwakarma[1]

[1]*Student, Electronics & Telecommunication Engineering, Mumbai University, Mumbai, Maharashtra, India*

Abstract - *The speedy growth in E-Commerce industry has led to an aggressive increase in the use of credit cards for online purchases. In recent years, credit card fraud has become a major complication for banks as it has become very difficult for detecting fraud in the credit card system. To overcome this obstacle Machine learning plays an important role in detecting the credit card fraud in the transactions. As in Machine learning the machine is trained and it predicts the output so, to predict the various bank transactions various machine learning algorithms are used. This paper examines the performance of K-nearest neighbors, Decision Tree, Logistic regression and Random forest for credit card fraud detection. The dataset considers fraud transactions as "Class 1" and valid ones as the "Class 0". The data set is highly imbalanced, it has about 0.173% of fraud transactions and the rest are valid transactions. Four different machine learning classification algorithms used here and the task is implemented in Python language. The performance of the algorithm is evaluated by accuracy score, f1-score, precision and recall score.*

Key Words: **Fraud Detection, Machine Learning, Logistic regression, K-nearest neighbors, Decision Tree, Random forest.**

1. INTRODUCTION

'Fraud' in credit card transactions is nothing but unauthorized and undesirable usage of an account by someone other than the owner of that account. Mandatory prevention measures can be taken to terminate this misuse and the behavior of such fraudulent practices can be studied to reduce it and to protect against similar occurrences in the future. In simple words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for some reasons while the owner and the card-issuing authorities are unfamiliar of the fact that the card is being used.

Credit Card Fraud is one of the massive threats to businesses today. However, to fight the fraud completely, it is important to first understand the structure of executing a fraud. Credit card fraudsters opt many numbers of ways to commit fraud. Card fraud is done either with the theft of the physical card or with the important information related with the account, including the information of the card account number or other information that necessarily be available to a practice a transaction.

Card numbers are often printed on the card, and a black stripe on the back contains the data in a machine-readable form. It contains the name of cardholder, card number, Expiration date, Verification/CVV code, type of card and more methods to commit credit card fraud.

A major challenge in applying Machine Learning to fraud detection is the presence of highly imbalanced dataset. In many available datasets, the majority of transactions are valid or genuine with an extremely small percentage of fraudulent ones. Designing an accurate fraud detection system that has low fraud transaction as compared to valid transactions hence detecting fraudulent activity effectively is a significant challenge for researchers. In our paper, we apply multiple classification approaches such as K-nearest neighbors, Decision Tree, Logistic regression and Random forest. Our aim is to build a classifier which will be able to separate fraud transactions from genuine ones. We will compare the accuracy and effectiveness of these algorithms in detecting fraud transactions.

2. LITERATURE REVIEW

A.Shen etal (2007)[1] demonstrate the efficiency of classification models to credit card fraud detection problem and the authors proposed the three classification models ie., decision tree, neural network and logistic regression. Among the three models neural network and logistic regression outperforms than the decision tree.

M.J.Islam et al (2007)[1] proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbor classifier is implemented and applied to the dataset for credit card system.

A Survey of Credit Card Fraud Detection Techniques [4] reviews and compares such multiple state of the art techniques, datasets and evaluation criteria applied to this problem. It discusses both supervised and unsupervised ML based approaches involving ANN (Artificial Neural Networks), SVM (Support Vector machines), HMM (Hidden Markov Models), clustering etc.

3. METHODOLOGY

3.1 Proposed Method

The proposed techniques focuses to detect the Credit Card Fraudulent on transactions whether it is a genuine or a fraud transaction and the approaches used to separate fraud and non-fraud are K-nearest neighbors, Decision Tree, Logistic regression and Random forest and Finally we will determine which approach is best for detecting credit card frauds. The figure below shows the system architecture diagram.

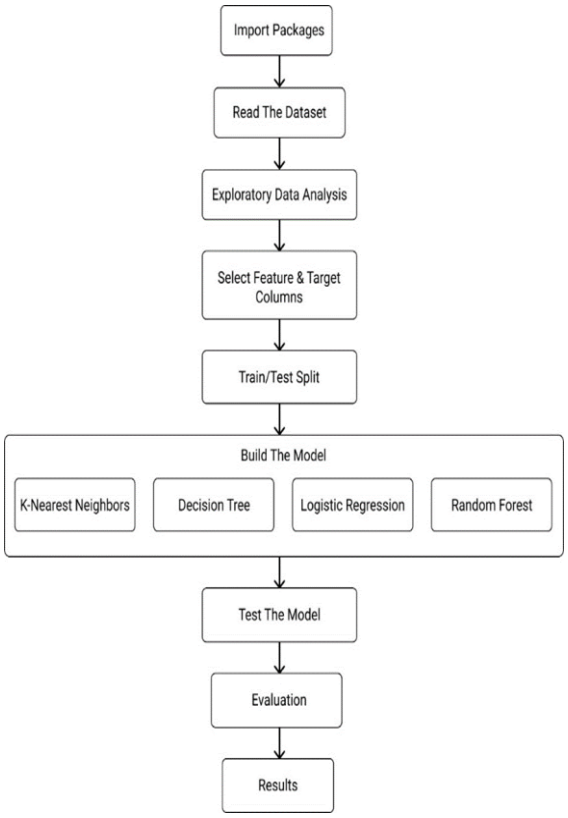

Fig -1: System Architecture

The system architecture has following steps:

1. Import Necessary Packages
2. Read the Dataset
3. Exploratory Data Analysis i.e. finding null values, duplicate values etc.
4. Selecting Features (X) and Target (y) columns
5. Train Test Split will split the whole dataset into train and test data
6. Build the model i.e. Training the model
7. Test the model i.e. Model prediction
8. Evaluation of the system i.e. Accuracy score, F1-score etc.

3.2 K-nearest Neighbors

The K-nearest neighbors algorithm is a classification algorithm that takes a bunch of labelled points and uses them to train the system on how to label other points. This algorithm classifies data points based on their similarity to other datapoints. In KNN, the data points that are nearest to each other are said to be Neighbors and similarly, in comparison we can say that the same class labels are near each other. Thus, there are different ways to measure the distance or dissimilarity of two data points such as Euclidean Distance or Minkowski etc.

K-nearest neighbors classifier implements learning based on the nearest neighbors of each query point, where 'k' is an integer value specified by the user. Radius Neighbors Classifier implements learning based on the number of neighbors within a fixed radius 'r' of each training point, where 'r' is a floating-point value specified by the user. The value of 'k' depends on various parameters but in our proposed methodology, the best value of 'k' is determined by a block of code.

The K-nearest neighbor algorithm

1. Pick a value for K.
2. Calculate the distance of unknown case from all cases.
3. Find the K observation in the training data that are nearest to the measurement of the unknown data point.
4. Predict the response of the unknown data point using the most popular value from the KNN.

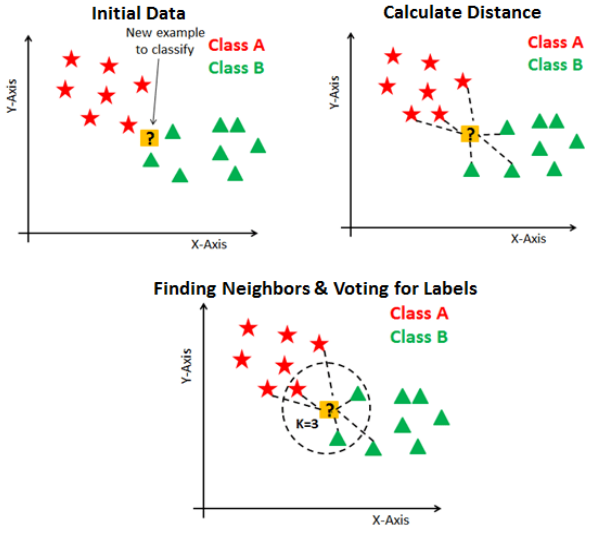

Fig -2: K-nearest Neighbor

So, from the above figure we can see that new example which is to be classified is set as a new example then the distance is calculated from each datapoints and increasing the value of K until the performance is satisfied. Finally, the distance which is closest to the example is assigned with that datapoint's class.

3.3 Decision Tree

A decision tree is built by splitting the training set into distinct nodes, where one node can contain all of or most of one category of the data. Decision Tree is built using recursive partitioning to classify the data. First, we select an attribute and this attribute should be the best attribute to split the data. The data should be split by minimizing the impurity at each step. Impurity of a node is calculated by the entropy of data in the node. Entropy is a measure of randomness or uncertainty, in simple words, Entropy of the node is how much random data is in that node. The lower the entropy the purer the node. There's one more definition Information Gain, it is the information that can increase the level of certainty after splitting.

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node is also called as terminal node represents the outcome and which do not split further. The top node in a decision tree is known as the root node. It tries to compute partition on the basis of the attribute value. It splits the tree in recursive manner called as recursive partitioning. This flowchart-like structure helps you in decision making.

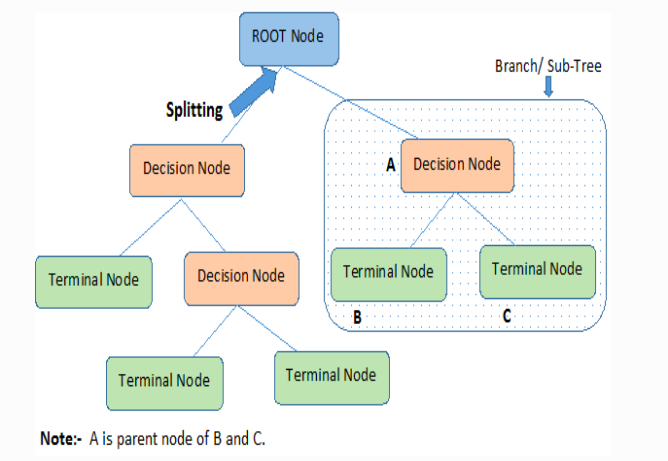

Fig -3: Decision Tree algorithm

1. **Root Node:** It represents the maximum population of the dataset and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing or splitting a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is said to be the decision node.
4. **Leaf / Terminal Node:** Nodes which do not split is called Leaf or Terminal node.
5. **Pruning:** Once we remove sub-nodes of a decision node, this process is termed as pruning. It is the opposite process of splitting.
6. **Branch / Sub-Tree:** A subsection of the complete tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is split into sub-nodes is said to be the parent node of sub-nodes whereas sub-nodes are the child of a parent node.

3.4 Logistic Regression

Logistic Regression is a classification algorithm for categorical values. In Logistic regression, we use one or more than one independent variables (X) to predict an outcome dependent variables (y). Logistic Regression is similar to Linear regression but tries to predict a categorical or discrete target field(label) instead of a numeric one. Such as yes/no, true/false, successful/unsuccessful, pregnant/not pregnant etc.

In Logistic regression independent variable(X) should be continuous, if categorical they must be dummied or indicator coded i.e. transform them into a numeric value. In Logistic Regression rather than fitting a straight line or a hyperplane, the logistic regression makes use of model uses the logistic function i.e. sigmoid function to extract the output of a linear equation which is between 0 and 1. The logistic function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1+\exp(-\eta)}$$

In the linear regression model, we have modelled the relationship between outcome and features with a linear equation:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

Therefore the finally expression is given as:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x).

3.5 Random Forest

Random forest is a tree-based algorithm which involves creating several trees and connecting with the output to reinforce ability of the model. Random forest is also a supervised learning algorithm. The term "forest" it's nothing but a set or bunch of decision trees.

Simply put, a random forest is built from numerous decision trees and helps to tackle the matter of overfitting in decision trees. These decision trees are randomly constructed by choosing random features from the given dataset. Random forest arrives at a call decision or prediction that has the maximum number of votes received from the decision trees. The end result which is found is nothing but the result which comes a maximum number of times through the numerous decision trees is taken into consideration as the final outcome by the random forest.

Random Forest is used to compute regression and classification problems. In regression problems, the dependent variable is continuous i.e. Numerical or Discrete values. In classification problems, the dependent variable is categorical i.e. Binary, true/false etc.

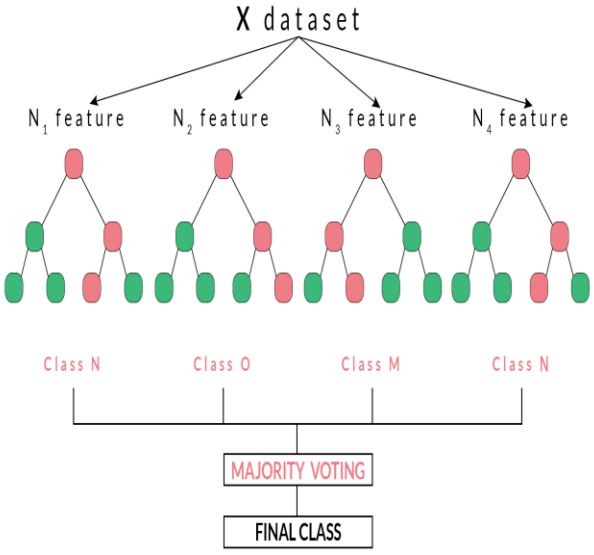

Fig -4: Random Forest Algorithm

For example, in the above diagram, we can see that each decision tree has voted or predicted a specific or a particular class. The final output or class selected by the Random Forest will be the Class N, because it has majority votes or is the predicted output by two out of the four decision trees.

4. EXPERIMENTAL RESULTS

4.1 Evaluation criteria

To evaluate the results of the classification algorithms there are various parameter such as Accuracy score, classification report, F1-score, confusion matrix etc.

1) Accuracy – Accuracy is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

2) Confusion Matrix - A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

3) Precision (Specificity)- It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

4) Recall (Sensitivity) - It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) F1- score - F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1].

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

4.2 Results

In this paper, four machine learning algorithms were used to detect the fraud in credit card system. To evaluate the algorithms, 80% of the dataset is used for training and 20% is used for testing. Accuracy, F1-score, precision, and recall score are used to evaluate the performance these four approaches.

As shown in the Table 1. The accuracy score for KNN, Decision tree, Logistic Regression and Random forest each algorithm is great. The accuracy is 0.99835, 0.99947, 0.99926 and 0.99958 respectively. But as we look to the other 3 criteria, we can clearly see that the **Random forest** classifiers outruns all the above classifier and predicts the fraudulent transaction with impressive F1 score, precision and recall score. Finally, the author visualizes these criteria in a graphical format as seen in Fig 5. which clearly represents that Random forest classifier is the best algorithm for credit card fraud detection.

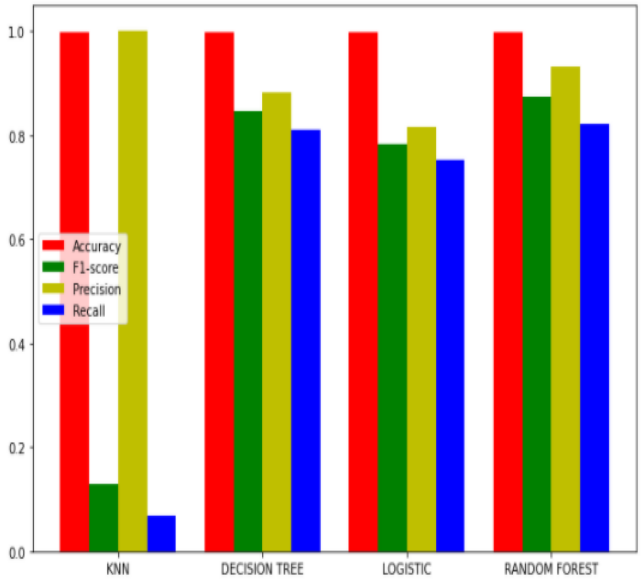

Fig -5: Graph

ALGORITHM	ACCURACY	F1-SCORE	PRECISION	RECALL
KNN	0.99835	0.12963	1.00000	0.06931
DECISION TREE	0.99947	0.84536	0.88172	0.81188
LOGISTIC REGRESSION	0.99926	0.78351	0.81720	0.75248
RANDOM FOREST	0.99958	0.87368	0.93258	0.82178

Table 1. Summary of Evaluation criteria

5. CONCLUSION

In this report, Machine learning technique like K-nearest neighbor, Decision tree, Logistic regression and Random forest were used to detect the fraud in the credit card system. Accuracy, F1-score, Confusion matrix, Sensitivity, Specificity were used to evaluate the performance of the proposed system. The accuracy for each and every algorithm was great but when other evaluation criteria comes into picture it was found that Random forest classifier is better than the KNN, Logistic regression and even Decision tree.

REFERENCES

[1] Machine Learning For Credit Card Fraud Detection System, Lakshmi S V S S, Selvani Deepthi Kavila, November 2018.

[2] Credit Card Fraud Detection using Machine Learning and Data Science, S P Maniraj, Aditya Saini , Shadab Ahmed, Swarna Deep Sarkar, September 2019.

[3] Fraud Detection using Machine Learning, Aditya Oza, 2019.

[4] A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective - Samaneh Sorournejad, Zojah, Atani et.al - November 2016.