IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 7, Issue 9, September 2020

DOI 10.17148/IARJSET.2020.7903

Credit Card Fraud Detection

M.Mohanapriya[1], M. Kalaimani[2]

Associate Professor, Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India[1]

Student, Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India[2]

Abstract: The main objective of this project is to develop a Credit card fraud Detection using the Random Forest Algorithm. Recently, a dramatic spike in the number of credit card purchases has contributed to a substantial uptick in fraudulent activity. Implementation of successful fraud prevention mechanisms has been necessary for all banks issuing credit cards to reduce their losses. This makes it difficult for the retailer to check whether or not the client who is making a transaction is the genuine cardholder. With the proposed method, the accuracy of detecting the fraud can be increased using random forest algorithm. Random forest algorithm classification method for study of data collection and actual consumer dataset. Finally optimize the precision of the data on the test. The techniques efficiency is judged based on accuracy, flexibility, specificity and precision. Then the analysis of some of the given attributes determines the identification of fraud and gives visualization of the graphical model. The performance of the techniques is measured based on precision, flexibility, specificity and accuracy.

Keywords: Credit card, Random Forest algorithm, Machine Learning, Decision Tree, and Classifier.

I. INTRODUCTION

In credit card purchases, numerous fraudulent behavior identification approaches have been applied in the minds of researchers including strategies for designing models focused on artificial intelligence, data processing, fuzzy logic, and machine learning. Detection of credit card fraud is extremely difficult, but also a common challenge to solve. Machine learning is known as a good tool for detecting fraud. During online transaction processes a significant amount of data is transmitted which results in a binary result: genuine or fake. Applications are built inside the fake sample datasets. There are data points, including the age and size of the customer account, and the credit card sources. There are hundreds of features which each adds to the risk of fraud, to varying degrees. Remember, the amount to which and function contributes to the fraud score is created by the machine's artificial intelligence that is powered by the training set but not by a fraud analyst. In the case of card fraud, however, if the use of cards to conduct fraud is known to be high, the fraud weighting of a transaction with a credit card would be similarly so. If this were to decrease though, the degree of participation would be similar. Simply render, without complex programming such as manual analysis, these models self-learn. Detection of credit card theft using Machine learning is performed using the classification and regression algorithms. We use supervised learning algorithms such as Random Forest Algorithms to detect online or offline fraud card transactions. Random forest is an evolved version of Decision tree. Random forest has higher performance and precision than the other algorithms of machine learning. Random forest attempts to reduce the above problem of association by selecting only a subsample of the space of the function at each break. In general, it attempts to de-correlate the trees and prune the trees by providing a stop criterion for splits of nodes.

1.1 Scope of the Project

We developed a protocol or model for detecting the fraud behavior in credit card transactions in this proposed project. This device provides much of the important features required to identify illegitimate and legal transactions. With technologies evolving, the nature and frequency of fraudulent transactions are difficult to trace. With the rise of machine learning, artificial intelligence, and other related information technology areas, it becomes possible to simplify the process and save some of the significant amount of effort being placed into identifying fraudulent credit card practices.

II. PRESENT FRAMEWORKS

In the current framework, studies on a case study involving the detection of credit card fraud, where data normalisation is implemented before Cluster Analysis and findings from the use of Cluster Analysis and Artificial Neural Networks on fraud detection have shown that neuronal inputs can be reduced by clustering attributes. And promising results can be achieved by the use of structured data and MLP training can be extended to data. The foundation of this study was unsupervised instruction. The aim of this paper was to discover new ways to diagnose fraud and to improve the precision of the findings. The data collection for this paper is focused on transactional real-life data from a major European company, and the secrecy of personal information in data. An algorithm's accuracy is about 50 percent. The purpose of this paper was to find an algorithm and lower the estimate of costs. The result obtained was 23 percent and Bayes' low risk algorithm was the one they noticed.

III. PROPOSED SYSTEM AND DESIGN

In proposed System, we are applying random forest algorithm for classification of the credit card dataset. Random Forest is a Classification and Regression algorithm. In short, it's a list of classifiers for the decision tree. Random forest has an advantage over the decision tree, since it corrects the practise of adapting their training set too well. A subset of the training set is sampled randomly such that each branch is trained and then a decision tree is constructed, then each node splits on a feature chosen from a random subset of the total feature set. In random forest, training is incredibly fast also for large data sets with many features and data instances and since each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting.

3.1 Advantages of Proposed System Over Existing System

- Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by Random Forest.
- The 'amount' feature is the transaction amount. Feature “class” is the target class for the binary classification and it takes value 1 for positive instances (fraud) and 0 for negative instances (not fraud).

3.2 Algorithm Used: Random Forest Algorithm:

Random Forest is a type of supervised algorithm for machine learning based on learning the ensemble. Ensemble learning is a method of learning where you multiply combine various types of algorithms or the same algorithm to create a more efficient model of prediction. The random forest algorithm incorporates multiple algorithms of the same form i.e. multiple decision trees, resulting in a trees forest, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

3.3 Working of Random Forest Algorithm:

The following are the basic steps involved in executing the algorithm for random forest
1. From the list select N random items.
2. Build a tree of decision dependent on certain records of N.
3. In your algorithm, pick the number of trees you like, and repeat steps 1 and 2.
4. Decision tree in the forest predicts the division to which the new record belongs, for classification.

3.4 Architecture

First, the credit card dataset is taken from the source, cleaning and validation is done on the dataset that involves eliminating duplication, filling empty spaces in columns, translating the appropriate component into variables or groups, and splitting the data into two sections, one is the testing dataset and another is the test data collection. The initial sample is now partitioned into the test and train data set.

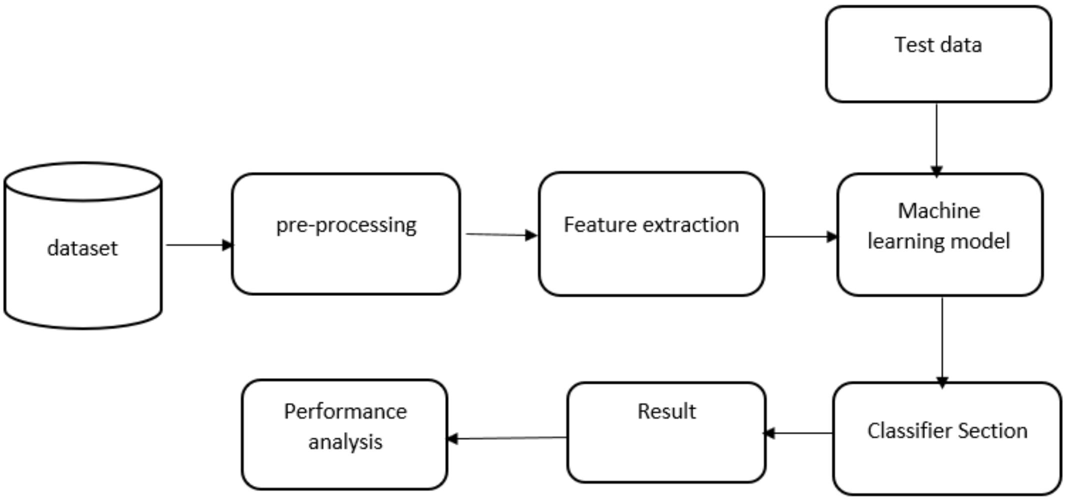

Fig.1 Block Diagram of the System

4.4 MODULES

A. MODULE 1: Data Collection

Data included in this paper is a compilation of product ratings obtained from databases of credit card purchases. This phase involves the specification of the subset of all available data for which you will be operating. ML problems ideally start with data, loads of data (examples or observations) that you already know for the target answer. Data for which the desired response is already identified is called labelled data.

B. MODULE 2: Data Pre-Processing

Organize your selected data by formatting, cleaning and sampling from it. Three common data pre-processing steps are:

- Formatting: The data you have chosen may not be in a format that suits you to work with. The data may be in a relational database, and you'd like it to be in a flat disc, or the data may be in a proprietary file format, whether you like it in a relational database or text file.
- Cleaning: Cleaning data involves removing or replacing lost data. Instances of data can be insufficient and do not hold the details that you think you need to fix the issue. May need to delete these instances. In addition, any of the attributes may have confidential details, and these attributes will need to be deleted completely from the data.
- Sampling: There could be much more data available to pick than you need to deal with. For algorithms and greater processing and memory needs, more data may result in much longer running times. Before evaluating the whole dataset, you should take a smaller representative sample of the collected data, which could be much easier to analyse and test solutions.

C. MODULE 3 : Feature Extraction

Next thing is to do extraction of Functionality is a method of attribute reduction. Unlike the collection of features which rank the current attributes according to their predictive importance, the extraction of features literally transforms the attributes. The converted attributes are linear variations of initial attributes, or functions. Finally, we practice our models using Classifier Algorithm . We use Python 's Natural Language Toolkit library identify feature. We use the extracted labelled dataset. The majority of our labelled data will be used for sample assessment. Some algorithms for machine learning were used to identify the pre-processed results. The classifiers selected were Random Trees. These algorithms are very common in tasks relating to text classification.

D. MODULE 4: Model Evaluation

Model Evaluation is an important part of the method of product development. It helps find the best model representing our results, and how well the model chosen will perform in the future. It is not appropriate in data science to assess model output with the data used for testing, since it can easily produce overoptimistic and over-fitted models. In data science, there are two methods of evaluating models, of evaluating model efficiency. -- classification model's performance is calculated based on its average. The outcome is in the shape that is visualized . Representation of graded data in graph form. The accuracy of the test data is known as the percentage of accurate predictions. It can be conveniently determined by calculating the number of predictions accurate by the number of predictions overall.

IV. SNAPSHOTS OF THE DESIGN

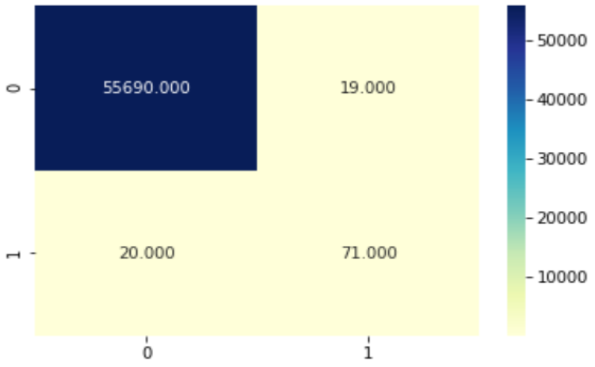

Fig 2. Classification of Data

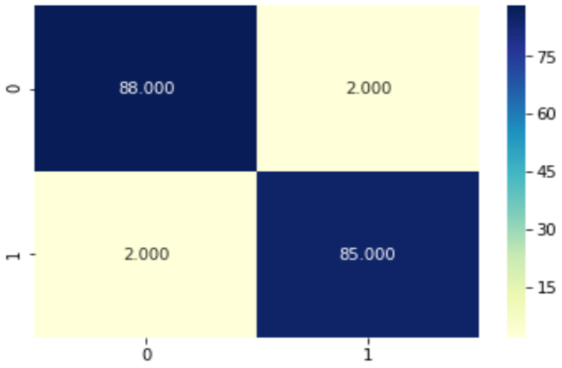

Fig 3. Classification of Data

DOI 10.17148/IARJSET.2020.7903

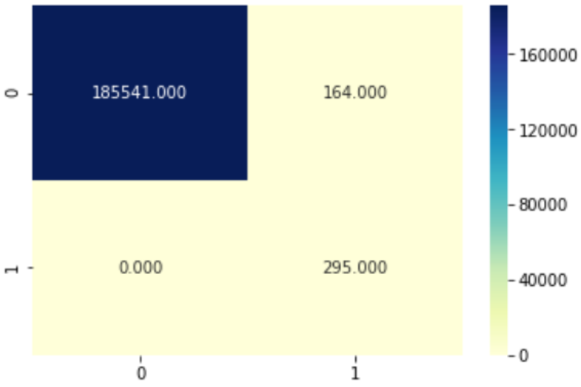

Fig 4. Classification of Data

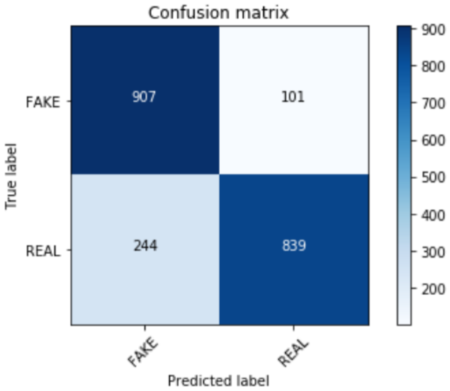

Fig 5. Accuracy with the Final Classification

V. CONCLUSION AND FUTURE WORK

With a greater number of training details, the Random Forest Algorithm will do better, but speed will suffer during testing and implementation. And it will benefit to add more pre-processing techniques. The SVM algorithm still suffers from the imbalanced data set issue and needs more pre - processing to provide better results on the results seen by SVM is fantastic but it should have been better if more pre-processing has been done on the data. Thus, the Random Forest Algorithm serves the best accuracy than any other machine learning algorithms.

REFERENCES

1. The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada (Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavol Zavarsky), June 2016.
2. BLAST-SSAHA Hybridization for Credit Card Fraud Detection, Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar, February, 2009.
3. Research on Credit Card Fraud Detection Model Based on Distance Sum, Wen-Fang YU, Na Wang, April, 2009.
4. Fraudulent Detection in Credit Card System Using SVM & Decision Tree, Vijayshree B. Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande, May, 2016.
5. Supervised Machine (SVM) Learning for Credit Card Fraud Detection, Sitaram Patel, Sunita Gond, February, 2014.
6. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Y. Sahin and E. Duman, March ,2011.
7. Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System, Dahee Choi and Kyungho Lee, Dec 2017.
8. Credit Card Fraud Detection Using Decision Tree Induction Algorithm, Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar, April, 2015.
9. Data Mining Techniques for Credit Card Fraud Detection: Empirical Study, Marwan Fahmi, Abeer Hamdy, Khaled Nagati, April, 2019.
10. Card Fraud Detection Using Learning Machine, Gheorghe Asachi" din Iaşi, February, 2019.