IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

Credit Card Fraud Detection using Random Forest

Lokesh Naik[1], K. Shekar[2], R Madhu[3], T. Vinod[4]
Assistant Professor, Dept. of Computer Science and Engineering[1,2,3,4,]
MLR Institute of Technology, Dundigal, India

Abstract: *The project undertaken is extremely engrossed on Credit Card Fraud Detection. The rise within the number of card transactions in every sector had been resulting in many fraudulent transactions. The aim is to get goods without paying, or to get unauthorized funds from an account. This way of implementing the efficient method of detecting the fraud credit cards has been the most crucial for every bank which provides the credit cards in order to reduce their loses. The crucial & foremost challenge in construction of business is that neither the card holder nor the card has to be present when some purchase is being made. So, this makes it impossible for the merchant to verify whether the customer is an authentic cardholder or not. With the proposed scheme, the accuracy of detecting the fraudulent transactions is improvised using random forest algorithm. The Classification process of random forest algorithm is to investigate data set and user current dataset. At last, the result obtained is further gone for optimizing the accuracy. And also, this is chosen to be the best as this technique's performance is best in accuracy, precision and thus can be easily evaluated and supported*

Keywords: Credit Card, Fraud Detection, Random Forest

I. INTRODUCTION

In today's technological world, we discover there are various fraudulent activities detection techniques which are implemented in card transactions are kept in researcher minds to methods to develop models supported various technologies like machine learning, data processing, computer science. It's not such easy to implement the credit card fraud detection, but it's also a well-liked problem to unravel. In the planned system, identifying & detecting the model is done with the use of Machine Learning. It's all known that everything is having its advancements following the technology and so is Machine Learning. It is been identified as the best field for fraud detection. During online transactions, an out sized amount of knowledge is transferred leading to a binary result: genuine or fraudulent. Now, we have few datasets which are fraudulent, so some features and patterns are been made. These features may include many factors such as the customer's age, value of the account etc. Thus, there are many features and aspects which can support in detecting the fraud.

In today's world, there are many people easily trying to commit the frauds and such can be prevented & reduced by this demo. We just need to prepare the models to self-learn making sure that there is no other external input. In this project, Credit Card Fraud Detection is done using Machine Learning, this follows the deployment of classification & regression algorithms. Though there are many algorithms in ML. We use a supervised learning algorithm known already, Random Forest, this is an algorithm which is best suitable for classifications. Here, we are using it as it can classify the transactions. As said, this is the next level of decision trees which has the best efficiency in order to result the best output with great accuracy. This includes many models which can solve many problems to provide efficient & accurate outputs.

1.2. Problem Definition

As we see in today's world, there are many billion dollars of loss that is occurred every year by these fraudulent card transactions. This fraud ca never be expected and may also take various forms of making it happen. An economic survey which has been conducted globally has put forward a statement that nearly 48% of the organizations has experienced economic crime. Thus, this should definitely be considered in order to reduce these frauds. Additionally,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

these trends in technology have become a plus point for the criminals to commit the fraud, & this usage of the cards has been rapidly increasing, so as is the crime rate for the fraudulent transactions. This effect of the fraud transactions is not only affecting the merchants, banks but also the common person using the credit card. In return, this might damage the reputation of the merchant as there will be huge financial losses, which are very difficult to clear out within a short span of time, this will affect few banks and holders in such a way that customers lose the trust & prefer some other way of making their credit cards.

1.2 Scope of the Project

In the project chosen, there is model (protocol) which is being designed for the detection of the fraud activity which are occurred in the credit card transactions.

And this is easily capable of giving few required & necessary features for the detection of fraudulent & legit transactions

As it is known to everyone about the technology & it is changing very rapidly, It usually becomes very difficult to trace the behavior and pattern of the occurring fraudulent transactions.

With the development in the world technologies, and also in few like AI & some related fields, It is becoming attainable to automate the method and it saves much number of effective amount of work that is been kept for the detection of the fraudulent activities. This is providing the best way to reduce the fraud detection.

II. RELATED WORK

[1] The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada. "KosemaniTemitayo Hafiz, Dr. Shaun Aghili, Dr. PavolZavarsky."

This research paper particularly is being focused on making the card from some related criteria which reflects the use & analysis used for detecting the credit card fraud. This states few of the resources from Canada & this uses the predictive analytic technology.

[2] Research on Credit Card Fraud Detection Model Based on Distance Sum., "Wen-Fang YU, Na Wang".

As said that there is a wide range in increase in frauds & easy way of attacking the systems of finance, which is also a part of the huge growth in the trade. Other side of the coin is the drastic & rapid change in the frauds which is drastically growing every day. This is taken in order to prevent the frauds & thus make up a model for this credit card fraud detection. This is using the outlier detection which can help in detecting the frauds at the same time, few statements show that the model is genuinely perfect & possible in detecting the fraud producing the accurate result.

[3] Supervised Machine (SVM) Learning for Credit Card Fraud Detection. "Sitaram Patel, Sunita Gond".

We know that there are methods for solving these detections, one such is the SVM (Support Vector Machine) based method which involves many kernels with multiple support and also accepts a lot from the user profile. This output is very simple & can be understood to an extent with its simulation result displaying the TN, FN along with variations improved in TP, FP rate.

[4] Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. "Y. Sahin and E. Duman"

There are various studies which state that classification models which are usually based on the decision trees and also the SVM (Support Vector Machines) are organized in order utilize them in correct format for this fraudulent detection. This includes the performance of SVM about how it is done & those decision trees help in using methods for the detection some real data set.

III. SYSTEM ANALYSIS

3.1 Existing System

In the previous systems, there are few case studies which have the statements involving the credit card fraud detection. It states that the cluster analysis is performed before which the data normalization is done and as per the output obtained from the provided analysis, As we know that there are many ways to solve, This Artificial Neural Network along with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

Cluster Analysis & can be very widely utilized on the fraudulent detection. We have to train the data in these systems called the MLP training. This research was supported unsupervised learning. The main & important aspect of the paper is seeking out new methods for the fraudulent detection & also make out new methods for making the best accuracy.

3.2 Detection of Fraud in Credit Card System using Decision Tree & SVM

The previous part of this existing system, can detect several activities which include this fraud detection & this with the rapid increase in every field as electronic commerce. This part of increasing the technology is good part of one side, where as the increase in fraud is causing huge debt to many companies. This is not just affecting the large sectors but a lot also to the individual people. As known, there are already many methods for detection of the credit card frauds like the Decision trees, HMM, Genetic algorithms etc. We also have AI from different concepts to solve these problems. In order to make all this a perfect stop, we use this algorithms & thereby reducing the huge loses.

A. Disadvantages

Here, As known there is a replacement collaborative method & to get to measure that something represents the gains and losses because of fraud detection is proposed.
And there exist something value sensitive way of method which is executed on Bayes minimum risk which is presented using the proposed cost measure.
An existing one has a phase where the cluster analysis is done after the data normalization & these results effect the attributes on the detection as it has to be minimized

3.3 Proposed Scheme

In the proposed system, For the classification of provided datasets of card, we are using the Random Forest algorithm. This is basically used for classification of the datasets & regression. As we all know, It's a set of decision tree classifiers. We can undoubtedly mention that this algorithm has great advantages over the many other algorithms as it helps in correct fitting. In this construction process, a training set is taken as a sample in order to train each one of it & such a tree is constructed. Though there are large data sets including various features & data, Random Forest helps in fast training as there is a property of training them individually. This has a great way for realizing the errors & make it best to over fitting.

A. Advantages of Proposed System

- Random forest is most important in identifying various variables in classification and regression problems and these are done using the Random Forest algorithm.
- There is a class for identifying different methods for classification, such is the amount & class which helps in easy understanding of the fraud which takes a positive case & the transaction which is not fraud takes zero case.

IV. REQUIREMENT SPECIFICATIONS

The requirements specification related to this project can be the specification especially the technical one, mentioning the software products. It's the primary step within the requirements analysis process it lists the necessities of a selected package including functional, performance and security requirements. the aim of software requirements specification is to supply a close overview of the software project, its parameters and goals.

4.1 Hardware Requirements

- Processor - Intel
- RAM - 4 Gb
- Hard Disk - 260 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

4.2 Software Requirements

- Python
- Google Collab
- OS - Windows 7, 8 and 10 (32 and 64 bit)

V. FEASIBILITY STUDY

5.1 Technical Feasibility

- It is obvious that required software & hardware are available for development and implementation for the proposed- system.
- It uses the software Anaconda.

5.2 Economical Feasibility

The cost for the detection proposed system is relatively and comparatively less to other existing software's.

5.3 Operational Feasibility

In this project, many required steps are need to be configured for the necessary software to work correctly.

VI. SYSTEM ARCHITECTURE

First In the beginning, the card which is to be checked has some steps in which the dataset is taken, then there will be cleaning & validation process performed. Here, hence there is automatic reduce to the redundancy & thereby can be named to as power as it is removing the repetency.

The next step is to split the data set into 2 parts, one named to be the training dataset & the another named to be the test data set. Finally, the sample is parted into the test & train dataset.

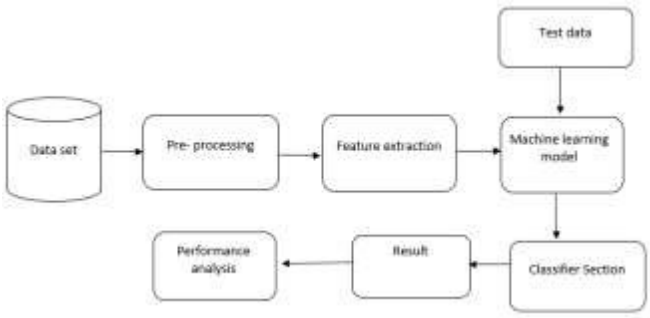


Figure 6.1: Architecture Of The Proposed System

VII. SYSTEM MODULES

7.1 Module Description

Module 1: Data Collection

Data employed in this paper could be a set of product reviews collected from card transactions records. This step is worried with selecting the subset of all available data that you just are going to be working with. ML problems start with data preferably, many data (examples or observations) that you already know the target answer. The *labelled data* is said to be the data where the data is known, target answer is also known

Module 2: Data Pre-Processing

By using the steps formatting, cleaning and sampling, we want to prepare the chosen data.

Three common data pre-processing steps are:

- Formatting: The information you've got selected might not be in an exceedingly format that's suitable for you to figure with. the info could also be in an exceedingly computer database and you'd prefer it in an exceedingly

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

file, or the info is also during a proprietary file format and you'd prefer it in a very electronic information service or a computer file.

- Cleaning: The fixing or the removal of the missing data is usually named to be cleaning process. There could also be data instances that are incomplete and don't carry the information you suspect you would like to deal with the matter. These instances might have to be removed. Additionally, there is also sensitive information in a number of the attributes and these attributes might have to be far from the information entirely.
- Sampling: There could also be much more selected data available than you would like to figure with. More data may end up in for much longer running times for algorithms and bigger computational and memory requirements. you'll be able to take a smaller sample distribution of the chosen data that will be much faster for exploring and prototyping solutions before considering the entire dataset.

Module 3: Feature Extraction

The following program is something to detect the attribute reducing named to be the feature extraction. Other than the present selection methods, which exist attributes by making them changed in feature extraction. These combinations of the real qualities can form some changes records, characteristics. Here, there is a use of required dataset & these models undergo the process of evaluation which uses the remaining data which is labelled data. In order to make sure that data is pre-processed, some methods related to machine learning will be applied & processed. This method here used the algorithm Random Forest as the classifier. These are widely used under the text classification task which help a lot in classifications.

Module 4: Evaluation Model

The model development process includes a step called model evaluation. It aids in the search for the most effective model that describes our data, as well as how effectively the chosen model will add in the long run. In the technological world today, especially in data science, this data which is used for the training is genuinely not acceptable for the evaluation model performance as it may cause to overfit models. The very basic two methods used for the data models testing are the cross-validation & hold-out & these both methods use a test set (not seen by the model) to evaluate model performance in order to minimize overfitting. Graphs are used to represent classified data

VIII. ALGORITHM UTILIZED

8.1 Random Forest

Random Forest is a kind of algorithm under the supervised machine learning algorithm which supports the ensemble learning. This way of ensemble learning is a kind of learning where we can join different types of algorithms together or the same algorithm repeatedly in order to form a stronger & powerful prediction model. This way of combining the algorithms of similar kind, known to be as multiple decision trees results in a forest of trees which is then named as & hence is the name “Random Forest”. The Random Forest algorithm is known to be utilized for classification, regression tasks.

A. Working of Random Forest

The steps required in performing & executing the Random Forest algorithm are as follows:

1. Pick few random rows or records from the sample dataset name it be some N random records.
2. Then, we need to build some tree which supports these N rows or records.
3. Now, we need to select & choose a few trees which we would like in our algorithm & then repeat the steps thus creating some pairs.
4. In these cases of classification problems, we should know that there will be each tree within the forest which predicts the category to which the new record or row belongs. Finally, in the same way, the new record is assigned to the category to which there will be a maximum value vote and then it gets finalized.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

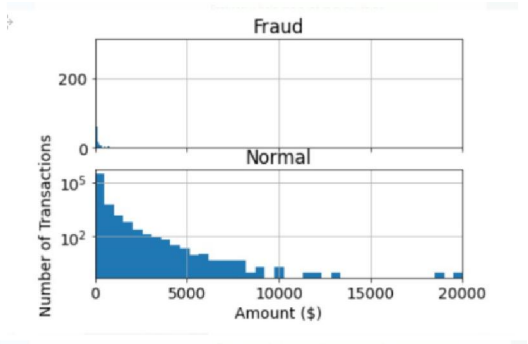
B. Advantages of Using Random Forest

There are few advantages or Pros of using random forest under classification and regression. They include:

1. As we already knew that the algorithm isn't biased, as there are many trees, each of it is trained on individually on a subset of data. Usually, this algorithm is based on the crowd, and hence the biasedness is minimum.
2. The next great pro is that this algorithm is extremely stable., even if an innovative information is introduced or not, the final algorithm will not be disturbed as the new data could depend o one tree, but it's not that possible to disturb or impact all the trees in the dataset.
3. The random forest algorithm works in a great way when you have got the features like the numerical & the categorical one, it is all set to work perfectly.
4. The random forest algorithm also works well when it has no proper data set which includes the missing data sets, not properly scaled one

IX. APPENDICES

9.1 Sample Screenshots from the Project



X. CONCLUSION

Random forest algorithm will perform better with a bigger number of coaching data, but speed during testing and application will suffer. And also, few Applications under the pre-processing techniques would also help this in a var. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to provide better results at the results shown by SVM is great but it could be better if more preprocessing is done on the info.

REFERENCES

[1] Sudhamathy G: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R, vol. 8, no-5, pp. 1954-1966.

[2] LI Changjian, HU Peng: Credit Risk Assessment for Ural Credit Cooperatives based on Improved Neural Network, International Conference on Smart Grid and Electrical Automation vol. 60, no. - 3, pp 227-230, 2017.

[3] Wei Sun, Chen-Guang Yang, Jian-Xun Qi: Credit Risk Assessment in Commercial Banks Based on Support Vector Machines, vol.6, pp 2430-2433, 2006.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

[4] Amlan Kundu, SuvasiniPanigrahi, Shamik Sural, Senior Member, IEEE, "BLAST-SSAHA Hybridization for Credit Card Fraud Detection", vol. 6, no. 4 pp. 309-315, 2009.
[5] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi Conference of Engineers and Computer Scientists, vol. I, 2011.
[6]Sitaram Patel, Sunita Gond, "Supervised Machine (SVM) Learning for Credit Card Fraud Detection, International of engineering trends and technology, vol. 8, no. -3, pp. 137- 140, 2014