INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Credit card fraud detection using Random Forest Algorithm

Vaishnave Jonnalagadda
jyvaishu@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Priya Gupta
priyagupta10061997@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Eesita Sen
eesita.sen@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

ABSTRACT

This Project is focused on credit card fraud detection in real-world scenarios. Nowadays credit card frauds are drastically increasing in number as compared to earlier times. Criminals are using fake identity and various technologies to trap the users and get the money out of them. Therefore, it is very essential to find a solution to these types of frauds. In this proposed project we designed a model to detect the fraud activity in credit card transactions. This system can provide most of the important features required to detect illegal and illicit transactions. As technology changes constantly, it is becoming difficult to track the behavior and pattern of criminal transactions. To come up with the solution one can make use of technologies with the increase of machine learning, artificial intelligence and other relevant fields of information technology, it becomes feasible to automate this process and to save some of the intensive amounts of labor that is put into detecting credit card fraud. Initially, we will collect the credit card usage data-set by users and classify it as trained and testing dataset using a random forest algorithm and decision trees. Using this feasible algorithm, we can analyze the larger data-set and user provided current data-set. Then augment the accuracy of the result data. Proceeded with the application of processing of some of the attributes provided which can find affected fraud detection in viewing the graphical model of data visualization. The performance of the techniques is gauged based on accuracy, sensitivity, and specificity, precision. The results is indicated concerning the best accuracy for Random Forest are unit 98.6% respectively.

Keywords— *Random forest algorithm, Criminal transactions, Credit card*

1. INTRODUCTION

Nowadays Credit card usage has been drastically increased across the world, now people believe in going cashless and are completely dependent on online transactions. The credit card has made the digital transaction easier and more accessible. A huge number of dollars of loss are caused every year by the criminal credit card transactions. Fraud is as old as mankind itself and can take an unlimited variety of different forms. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore, there's positively a necessity to unravel the matter of credit card fraud detection. Moreover, the growth of new technologies provides supplementary ways in which criminals may commit a scam. The use of credit cards is predominant in modern day society and credit card fraud has been kept on increasing in recent years. Huge Financial losses have been fraudulent effects on not only merchants and banks but also the individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that. For example, if a cardholder is a victim of fraud with a certain company, he may no longer trust their business and choose a competitor.

Fraud Detection is the process of monitoring the transaction behavior of a cardholder to detect whether an incoming transaction is authentic and authorized or not otherwise it will be detected as illicit. In a planned system, we are applying the random forest algorithm for classifying the credit card dataset. Random Forest is an associate in the nursing algorithmic program for classification and regression. Hence, it is a collection of decision tree classifiers. The random forest has an advantage over the decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature designated from a random subset of the complete feature set. Even for large data sets with many features and data instances, training is extremely fast in the random forest and because 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 overfitting.

1.1 Advantages

- Random Forest selects the best feature rather than the most important feature among a random subset of data resulting in a better model.
- Thus having a binary classification of fraud i.e. positive case (value 1) and non-fraud i.e. negative case (value 0) for the target category in the transaction amount.

There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on

artificial intelligence, data mining, fuzzy logic and machine learning. Credit card fraud detection is a very troublesome, but also a popular problem to solve. In our proposed system we built the credit card fraud detection using Machine learning. With the advancement of machine learning techniques. Machine learning has been recognized as a no-hit live for fraud detection. A great deal of data is transferred throughout on-line transaction processes, resulting in a binary result: genuine or fraudulent. Online businesses are able to identify fraudulent transactions accurately because they receive chargebacks on them. Within the sample fraudulent datasets, features are constructed.

These area unit information points like the age and price of the client account, as well as the origin of the credit card. There are many options and everyone contributes, to varying extents, towards the fraud probability.

Note, the degree within which every feature contributes to the fraud score isn't determined by a fraud analyst, but is generated by the artificial intelligence of the machine which is driven by the training set. So, in regard to the card fraud, if the use of cards to commit fraud is proven to be high, the fraud weighting of a transaction that uses a credit card will be equally so. However, if this were to diminish, the contribution level would parallel. Simply put, these models self-learn while not express programming like with manual review.

Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms. We use a supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction online or by offline. Random forest is an advanced version of the Decision tree. The random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by choosing only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by setting a stopping criterion for node splits

2. SOFTWARE AND HARDWARE REQUIREMENT

2.1 Hardware

- OS – Windows 7, 8 and 10 (32 and 64 bit)
- RAM – 4GB

2.2 Software

- Python
- Anaconda

3. SYSTEM ARCHITECTURE

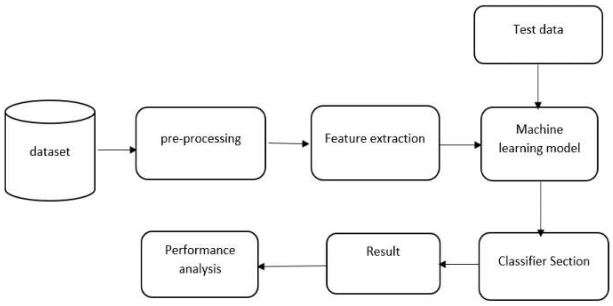

Fig. 1: System architecture

4. LITERATURE SURVEY

Fraudulent Detection in Credit Card System Using SVM & Decision Tree (Vijayshree B. Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande): With growing advancement in the electronic commerce field, fraud is spreading all over the world, causing major financial losses. In the current scenario, Major cause of financial losses is credit card fraud; it not only affects tradesperson but also individual clients. Decision tree, Genetic algorithm, Meta-learning strategy, neural network, HMM are the presented methods used to detect credit card frauds. In contemplating system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem. Thus by the implementation of this hybrid approach, financial losses can be reduced to greater extent.

Machine Learning Based Approach to Financial Fraud Detection Process in Mobile Payment System (Dahee Choi and Kyungho Lee): Mobile payment fraud is the unauthorized use of mobile transaction through identity theft or credit card stealing to fraudulently obtain money. Mobile payment fraud is a fast-growing issue through the emergence of smartphone and online transition services. In the real world, a highly accurate process in mobile payment fraud detection is needed since financial fraud causes financial loss. Therefore, our approach proposed the overall process of detecting mobile payment fraud based on machine learning, supervised and unsupervised method to detect fraud and process large amounts of financial data. Moreover, our approach performed sampling process and feature selection process for fast processing with large volumes of transaction data and to achieve high accuracy in mobile payment detection. F-measure and ROC curve are used to validate our proposed model.

5. PURPOSE OF THE PROJECT

We propose a Machine learning model to detect fraudulent credit card activities in online financial transactions. Analyzing fake transactions manually is impracticable due to vast amounts of data and its complexity. However, adequately given informative features, could make it is possible using Machine Learning. This hypothesis will be explored in the project.

To classify fraudulent and legitimate credit card transaction by supervised learning Algorithm such as Random forest. To help us to get awareness about the fraudulent and without loss of any financially.

5.1 Packages

Which are being used for data exploration, pro processing and for using random forest algorithm are:

- **NumPy**: For simple arrays.
- **Pandas**: For reading the file.
- **SciKit**: Learn- for pre-processing.
- **Matplotlib or Seaborn**: For plotting and representing confusion matrix colour format.
- **Tensor flow**: For matrix format.

6. MODULES

- Data collection
- Data pre-processing
- Feature extraction
- Evaluation model

6.1 Data collection

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

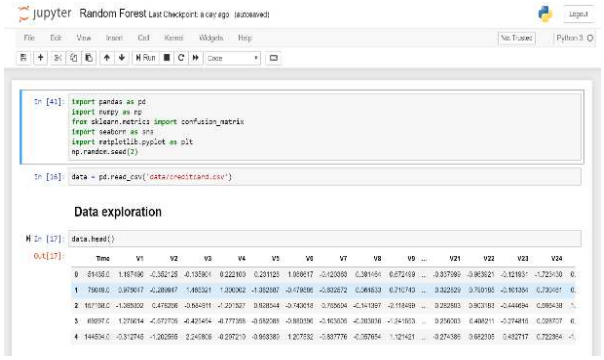

Fig. 2: Importing python packages for data exploration, preprocessing and for using random forest algorithm

6.2 Data pre-processing

Pre-processing is the process of three important and common steps as follows:

- **Formatting**: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is .csv files.
- **Cleaning**: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists, Data Cleaning continues of 80% of work.
- **Sampling**: This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behavior and pattern of data in an integrated way

6.3 Data exploration

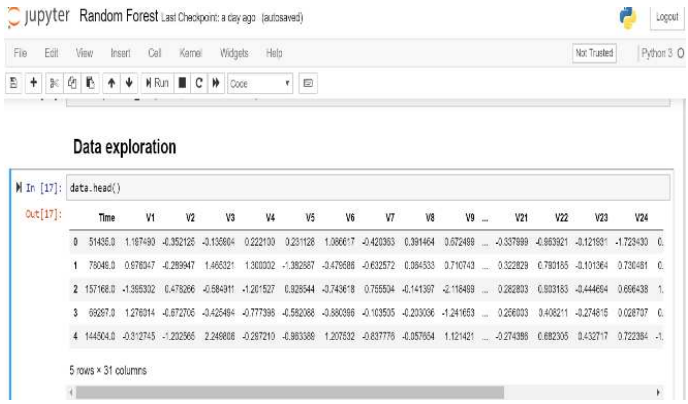

Fig. 3: Data exploration

6.3.1 Pre-processing with python commands

Step 1:

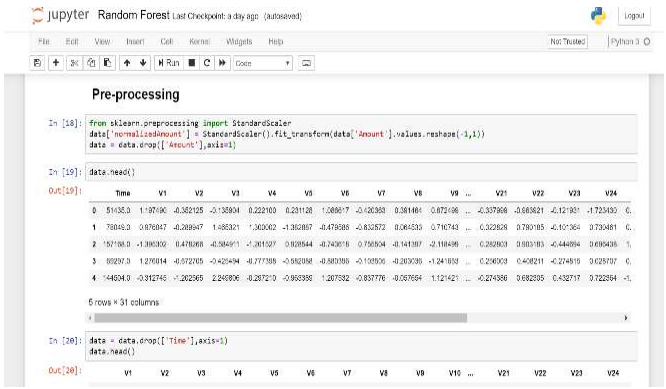

Fig. 4: Pre processing

Step 2:

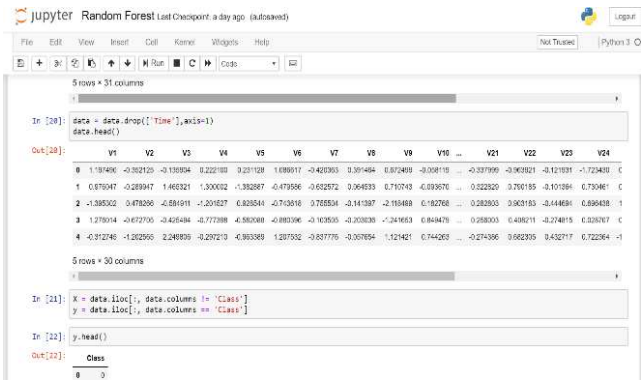

Fig. 5: Preprocessing Step 2

Step 3: Acquired trained and testing dataset from the large dataset

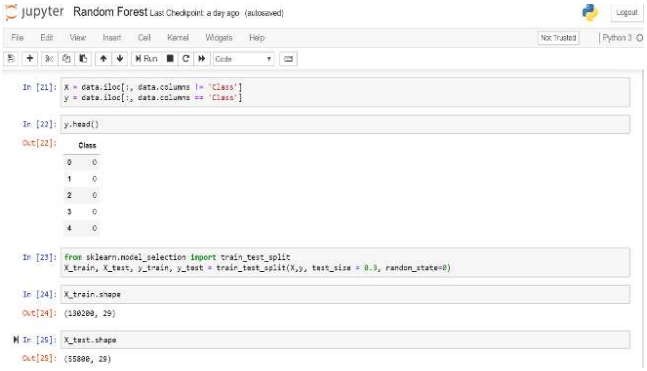

Fig. 6: Training and testing data

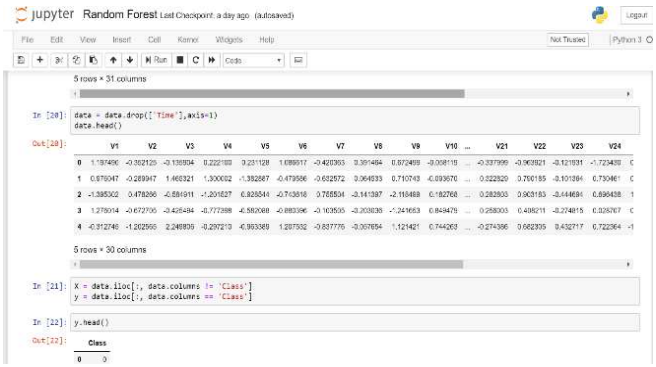

Fig. 7: Process of training and testing data extraction

6.4 Data visualization

Data Visualisation is the method of representing the data in a graphical and pictorial way, data scientists depict a story by the results they derive from analysing and visualising the data. The best tool used is Tableau which has many features to play around with data and fetch wonderful results.

6.5 Feature extraction

Feature extraction is the process of studying the behavior and pattern of the analyzed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

6.6 Evaluation model

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

7. ALGORITHM

7.1 Random forest

Random forest is a supervised machine learning algorithm based on ensemble learning. Ensemble learning is an algorithm where the predictions are derived by assembling or bagging different models or similar model multiple times. The random forest algorithm works in a similar way and uses multiple algorithm i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

7.1.1 Advantages of using random forest

- The random forest algorithm is not biased and depends on multiple trees where each tree is trained separately based on the data, therefore biasedness is reduced overall.
- It's a very stable algorithm. Even if a new data point is introduced in the dataset it doesn't affect the overall algorithm rather affect the only a single tree.
- It works well when one has both categorical and numerical features.
- The random forest algorithm also works well when data possess missing values, or when it's not been scaled properly.

Thus, using this Random forest algorithm and decision trees algorithm we have extracted the accurate percentage of detection of fraud from the given dataset by studying its behavior.

A confusion matrix is basically a summary of prediction results or a table which is used to describe the performance of the classifier on a set of test data where true values are known. It provides visualization of an algorithm's performance and allows easy identification of classes. Thus, resulting in the computing of most performance measures by giving insights not only the errors being made by the classification model but also tells the type of errors being made.

Trained Data and Testing Data is represented in a confusion matrix which portrays:

- **TP:** True Positive which denotes the real data where customers are subjected to fraud and are used for training and were accurately predicted.
- **TN:** True Negative denotes the data which was not predicted and doesn't match with the data which was subjected to the fraud.
- **FP:** False Positive is predicted but there is no possibility of the data to be subjected to the fraud.
- **FN:** False Negative is not predicted but there is an actual possibility of the data who is subjected to fraud.

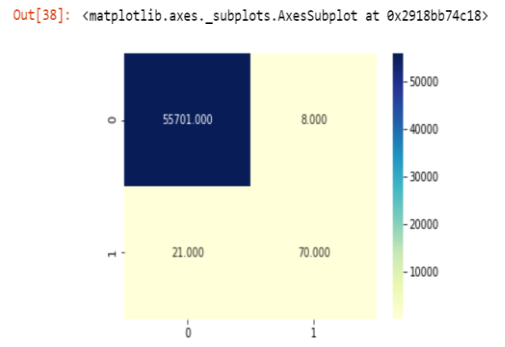

Fig. 8 Confusion matrix for testing dataset

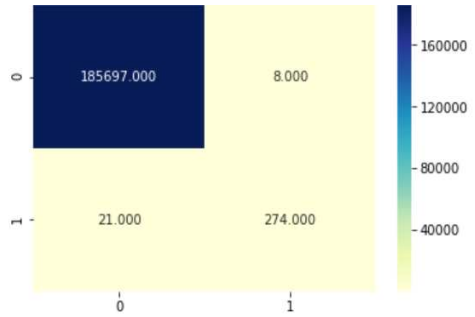

Fig. 9: Confusion matrix for testing dataset

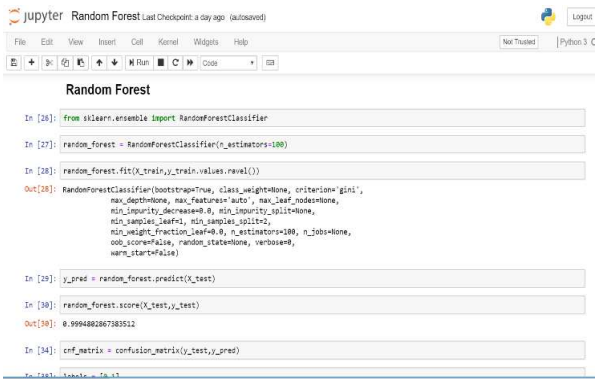

Fig. 10: Accurate result extracted from the random forest classification and regression model using decision trees

8. CONCLUSION

Hence, we have acquired the result of an accurate value of credit card fraud detection i.e. 0.9994802867383512 (99.93%) using a random forest algorithm with new enhancements. In comparison to existing modules, this proposed module is applicable for the larger dataset and provides more accurate results. The Random forest algorithm will provide better performance with many training data, but speed during testing and application will still suffer. Usage of more pre-processing techniques would also assist. Our future work will try to represent this into a software application and provide a solution for credit card fraud using the new technologies like Machine Learning, Artificial Intelligence and Deep Learning.

9. REFERENCES

[1] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[2] https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

[3] Gupta, Shalini, and R. Johari. "A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant." International

Conference on Communication Systems and Network Technologies IEEE, 2011:22-26.

[4] Y. Gmbh and K. G. Co, "Global online payment methods: the Full year 2016," Tech. Rep., 3 2016.

[5] Bolton, Richard J., and J. H. David."Unsupervised Profiling Methodsfor Fraud Detection." Proc Credit Scoring and Credit Control VII (2001):5– 7.

[6] Drummond, C., and Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. Proc of the ICML Workshop on Learning from Imbalanced Datasets II, 1–8. Quah, J. T. S., and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732.

