International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

Credit Card Fraud Detection Using Machine Learning

Maimuna Begum[1], Mohammadi Fatima[2], Sadaf Naaz Farooqui[3], Rafath Samrin[4],
[1,2,3]UG Scholar, Dept. of IT, ISL Engineering College, Hyderabad,
[4]Associate Professor, Dept. of IT, ISL Engineering College, Hyderabad.

ABSTRACT

In our project, in the main targeted on credit card fraud detection for in globe. at the start I'll collect the credit card knowledge sets for trained knowledge set. Then can give the user credit card queries for testing knowledge set. once classification method of random forest algorithmic program victimisation to the already analysing knowledge set and user give current knowledge set. Finally optimizing the accuracy of the result knowledge. Then can apply the process of a number of the attributes provided will notice affected fraud detection in viewing the graphical model visual image. The performance of the techniques is evaluated supported accuracy, sensitivity, and specificity, precision. The results indicate regarding the best accuracy for Random Forest are 98.6% severally.

Key Words: Fraud in credit card, data mining, logistic regression, decision tree, SVM, random forest, collative analysis.

INTRODUCTION

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. There are many fraud detection solutions and software which prevent frauds in businesses such as credit card, retail, e-commerce, insurance, and industries. Data mining technique is one notable and popular methods used in solving credit fraud detection problem. It is impossible to be sheer certain about the true intention and rightfulness behind an application or transaction. In reality, to seek out possible evidences of fraud from the available data using mathematical algorithms is the best effective option. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set mining, machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree and random forest is carried out. Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to issue of having only a limited amount of data, credit card makes it challenging to match a pattern for dataset. Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behaviour. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and

the results of researches are often hidden and censored, making the results in accessible and due to this itis challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes a limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviours always different that is the legit transaction in the past may be a fraud in present or vice versa. This paper evaluates four advanced data mining approaches, Decision tree, support vector machines, Logistic regression and random forest sand then a collative comparison is made to evaluate that which model performed best. Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature (variables) selection for the models, suitable metrics most important part of data mining to evaluate performance of techniques on skewed credit card fraud data. A number of challenges are associated with credit card detection, namely fraudulent behaviour profile is dynamic, that is fraudulent transactions tend to look like legitimate ones, Credit card fraud detection performance is greatly affected by type of sampling approach used, selection of variables and detection technique used. In the end of this paper, conclusions about results of classifier evaluative testing are made and collated.

Related Work

In This paper represents a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

A new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to23%is obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential., accuracy of an algorithm is around50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk. Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Survey of various techniques used in credit card fraud detection mechanisms has been Shown in this paper along with evaluation of each methodology based on certain design criteria. Analysis on Credit

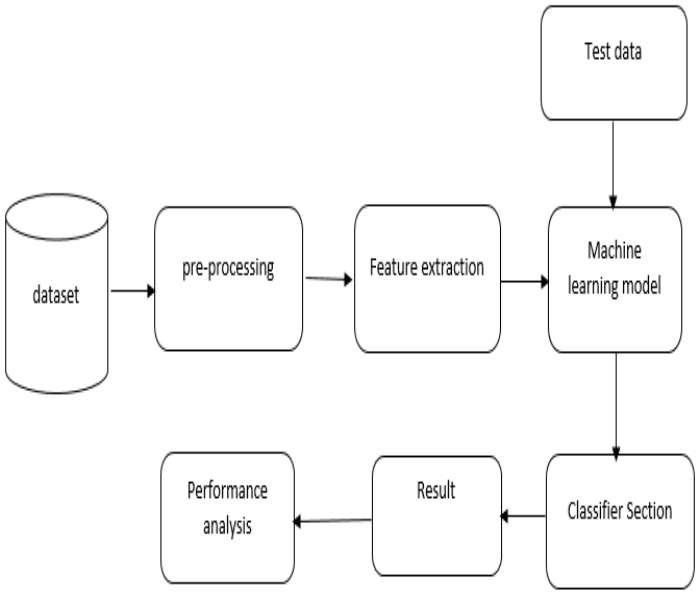
Card Fraud Detection Methods has been done. The survey in this paper was purely based to detect the efficiency and transparency of each method. Significance of this paper was conduct a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem.

EXPERIMENTAL SET UP AND METHODS

This section describes the dataset used in the experiments and the three classifiers under study, namely; Naïve Bayes, k-Nearest Neighbour and Logistic Regression techniques. The different stages involved in generating the classifiers include; collection of data, pre-processing of data, analysis of data, training of the classifier algorithm and testing (evaluation). During the pre-processing stage, the data is converted into useable format fit and sampled. A hybrid of under-sampling (the negative cases) and over-sampling (the positive cases) is carried out to achieve two sets of data distributions. For the analysis stage, the feature selection and reduction is already carried out on the dataset using PCA. The training stage is where the classifier algorithms are developed and fed with the processed data. The experiments are evaluated using True positive, True Negative, False Positive and False Negative rates metric. The performance comparison of the classifiers is analysed based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and balanced classification rate.

Dataset The dataset is sourced from ULB Machine Learning Group and description is found in [32]. The dataset contains credit card transactions made by European cardholders in September 2013. This dataset presents transactions that occurred in two days, consisting of 284,807 transactions. The positive class (fraud cases) make up 0.172% of the transactions data. The dataset is highly unbalanced and skewed towards the positive class. It contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. Thus a total of 30 input features are utilized in this study. The details and background information of the features cannot be presented due to confidentiality issues. The time feature contains the seconds elapsed between each transaction and the first transaction in the dataset. The 'amount' feature is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (no fraud).

SYSTEM ARCHITECTURE



1. DATA COLLECTION

Data used in this paper is a set of product reviews collected from credit card

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

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*.

2. DATA PRE-PROCESSING

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

3. FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using 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.

4. EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test

data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

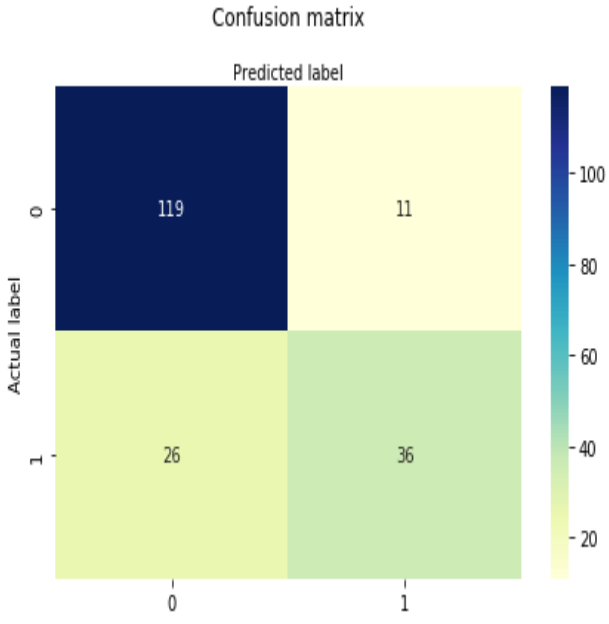
Result

Accuracy Result:

Models / No. of runs	Run 1	Run 2	Run 3
Decision Tree	0.85	0.93	0.88
Random Forest	0.92	0.99	0.98

CONFUSION MATRIX FORMAT

Actual/Predicted	Not a fraud	Fraud
Not a Fraud	True Positive	False Positive
Fraud	False Negative	True Negative



CONCLUSION

From the experiments the result that has been concluded is that Logistic regression has an accuracy of 97.7% while SVM shows accuracy of 97.5%and Decision tree shows accuracy of 95.5%but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6%in problem of credit card fraud detection with dataset provided by ULB machine learning. The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more pre-processing to give better results at the results shown by SVM is great but it could have been better if more pre-processing have been done on the data.

REFERENCES

[1]Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., (2002). Credit card fraud detection using Bayesian and neural networks.Proceeding International NAISO Congress on Neuro Fuzzy Technologies.
[2]Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311 – 322
[3]RamaKalyani, K. and UmaDevi, D., (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm, International Journal of Scientific & Engineering Research, Vol. 3, Issue 7, pp. 1 – 6, ISSN 2229-5518
[4]Meshram, P. L., and Bhanarkar, P., (2012). Credit and ATM Card Fraud Detection Using Genetic Approach, International Journal of Engineering

Research & Technology (IJERT), Vol. 1 Issue 10, pp. 1 – 5, ISSN: 2278-0181
[5]Singh, G., Gupta, R., Rastogi, A., Chandel, M. D. S., and Riyaz, A., (2012). A Machine Learning Approach for Detection of Fraud based on SVM, International Journal of Scientific Engineering and Technology, Volume No.1, Issue No.3, pp. 194-198, ISSN : 2277-1581
[6]Seeja, K. R., and Zareapoor, M., (2014). FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, Article ID 252797, pp. 1 – 10, http://dx.doi.org/10.1155/2014/252797
[7]Patil, S., Somavanshi, H., Gaikwad, J., Deshmane, A., and Badgujar, R., (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4, Issue 4, pp. 92-95, ISSN: 2320-088X
[8]Duman, E., Buyukkaya, A., & Elikucuk, I. (2013). A novel and successful credit card fraud detection system implemented in a turkish bank. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on (pp. 162-171). IEEE.
[9]Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 677-685). Society for Industrial and Applied Mathematics.
[10]Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2, 841-848. [11]Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).
[12]Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In Service Systems and Service Management, 2007 International Conference on (pp. 1-4). IEEE.