EXAMINING THE IDENTIFICATION AND DETECTION OF CREDIT CARD FRAUD WITH THE USE OF RANDOM FOREST ALGORITHMS

[1]K SHOBHA RANI, [2]V ANITHA, [3]DATTATREYA GOUDAPPOLA, [4]BANDLA SANDEEP
[1,2,3]Assistant Professor, [4]Student, Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT
Credit risk is one of the main functions of banking. Banks classify risk according to their profile. Although many algorithms came into existence still the issue is yet to solve. In existence, data normalization is applied before Cluster Analysis and the obtained results from Cluster Analysis and Artificial Neural Networks on fraud detection has shown by clustering attributes and the neuronal inputs can be minimized. Significance of the paper is to find an algorithm to reduce the cost measure. The result obtained was 23% and the algorithm used was Minimum Bayesian-Risk (MBR). In proposed system, Random Forest Algorithm is used for classification and regression. Random forest has the advantage over decision tree as it corrects the habit of over fitting to their training data sets. It has been found to provide a good estimate of generalization error and resistant to over fitting. In credit card fraud detection, credit card data sets are collected for trained data sets and user credit card queries are collected for testing data sets. After classification process, Random Forest Algorithm is used for analyzing data sets and current data sets. Finally, the optimization is done and the accuracy obtained by Random Forest is 99.9%.
Keywords: Credit card fraud detection, Random forests, Minimum Bayesian-Risk (MBR)

1. INTRODUCTION

Billions of losses are caused every year by the fraudulent credit card transactions. Fraud is old as humanity 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 [3]. Therefore, there's positively a requirement to resolve the matter of credit card fraud detection. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years [2]. Hugh financial losses have been fraudulent affects not only merchants and banks, but also individual person who is using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that, though difficult to quantify in the short term, may become visible in the long period [4]. For example, if a cardholder is victim of fraud with a precise company, he might no longer trust theirbusiness and opt for a rival.

2. EXISTINGSYSTEM

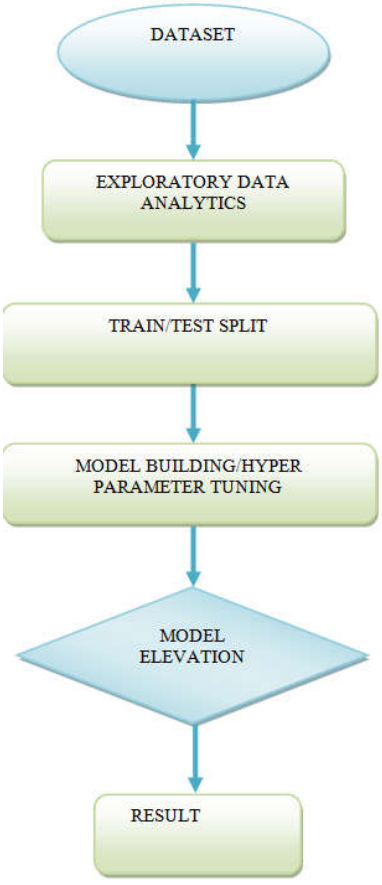
In existing System, a research about a case study involving credit card fraud detection, where data normalization is applied before Naïve Bayer's and Cluster Analysis and with results obtained from the use of these methods on fraud detection has shown that by clustering attributes neuronal inputs can be minimized and promising results can be obtained by using normalized data. This research was based on unsupervised learning. Significancedetection and to increase the accuracy ofresults. 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 around 50%.

Disadvantages

- Thus, the accuracy of the results obtained from these methods are less when compared with the proposed system.
- Less accuracy value.
- Less efficiency and security.

3. PROPOSED SYSTEM

The proposed System uses Random Forest Algorithm for classify the credit card data set. Random Forest is an algorithmic program for classification and regression. Summarily, it is a set of decision tree classifiers. Random Forest has advantage over decision tree because 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 build, each node then splits on a feature selected from a random subset of the total feature set. Even for large data sets with many features and data instances training is extremely fast in Random Forest and because each tree is trained independently of the others. The Random Forest Algorithm has been found to produce a good estimate of the generalization error and to be resistant to over fitting.



3.1 Advantages

- By applying the Random Forest Algorithm, the dataset will be classified into four categories which will be obtained in the form of confusion matrix.
- Based on the above classification of data performance analysis will be done.
- In this analysis the accuracy of credit card fraud transactions can be obtained which will be finally represented in the form of graphical representation.

cr

- Thus, the accuracy of the results obtained from the methods are high when compared with the Existing system.
- High accuracy, efficiency and security.

4. SYSTEM ARCHITECTURE

The architecture of Credit card fraud detection involves mainly 5 steps:

1. EXPLORATORY DATA ANALYTICS
2. TRAIN/TEST SPLIT
3. MODEL BUILDING/HYPER PARAMETER TUNING
4. MODEL EVALUTION
5. RESULT

The architecture of the system is show below:

5. PROJECT MODULES

Our proposed system is categorized into five modules. They arerepresenedas follows:

5.1Module –I:DATA UNDERSTANDING

Here, we need to load the data and understand the features present in it. This would help us choose the features that we will need for your final model. The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

5.2 Module– II:EXPLORATORY DATA ANALYTICS (EDA):

Normally, in this step, we need to perform univariate and bivariate analyses of the data, followed by featuretransformations, if necessary. For the current data set, because Gaussian variables are used, we do not need to perform Z-scaling. However, you can check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.

5.3 Module– III:TRAIN/TEST SPLIT

Now we are familiar with the train/test split, which we can perform in order to check the performance of our models with unseen data. Here, for validation, we can use the k-fold cross-validation method. We need to choose an appropriate k value so that the minority class is correctly represented in the test folds.

5.4 Module– IV:MODEL-BUILDING

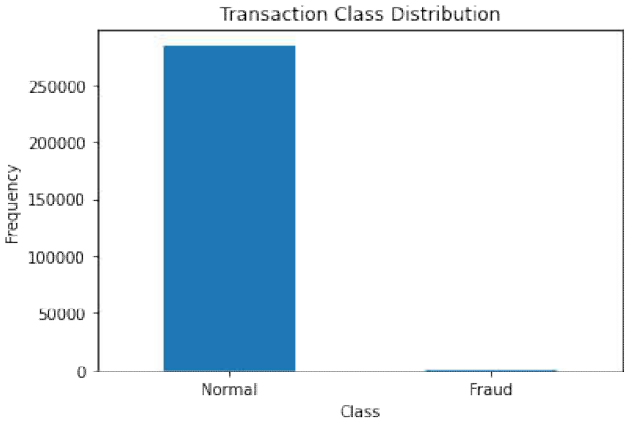
The main inspiration behind this type of learning is to learn from the information about the task, which has been provided in the past. A machine requires the basic data about the task to be provided to it. This basic input, or experience is given to it in the form of ‘training data’. This is the past information or data of a particular task.

5.5 Module– V:MODELEVALUATION

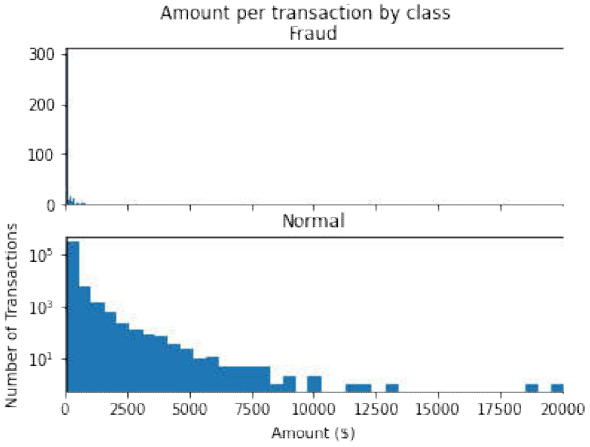
We need to evaluate the models using appropriate evaluation metrics. Note that since the data is imbalanced it is is more important to identify which are fraudulent transactions accurately than the non-fraudulent. We need to choose an appropriate evaluation metric which reflects this business goal.

cr

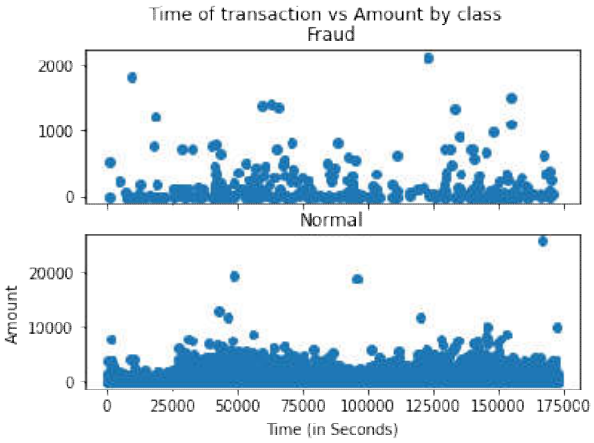
Screenshots



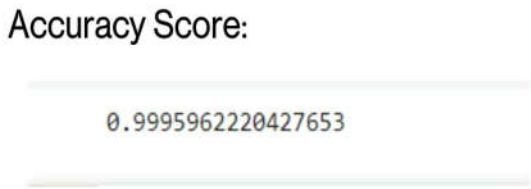
Output 1: Data set in Bar graph



Output 2: Data set in Histogram



cr



Output 4: Accuracy

Output 5: Prediction

6. **Methods**

- **Random forest**:Random forest is a basically supervised learning algorithm that is used for both classifications as well as regression.
- Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better the solution for the system.
- For getting high accuracy we used the Random Forest algorithm which gives accuracy which predicate by model and actual outcome of predication in the dataset.
- In the random forest which crates the decision tree from a sample of data and trees gives the prediction from each family and selects the best solution by voting which gives better accuracy for the model. It gives optimum results for the system.

CONCLUSION

In this study, we used an imbalanced dataset to check the suitability of different supervised machine learning models to predict the chances of occurrence of a fraudulent transaction. We used sensitivity, precision and time as the deciding parameters to come to a particular conclusion. Accuracy as a parameter was not used as it is not sensitive to imbalanced data and does not give a conclusive answer. We analyzed the kNN, Naive Bayes, Decision Tree, Logistic Regression and Random Forest models in this study. The Random Forest Algorithm will perform better with a larger number of training data, andthe result obtained is 99.9%. The SVM algorithm can be used instead of Random Forest, but it still suffers from the
imbalanced data set problem and requires more pre-processing to give better results.

FUTURE SCOPE

In future, privacy preserving techniques can be applied in distributed environment which will resolve the security related issues preventing private data access.This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and Logistic Regression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution. The prediction model predicts continuous valued functions. We have to

cr

detect 148 may be fraud and other are genuine. In decision tree generate a tree with root node, decision node and leaf nodes. The leaf node may be 1 becomes fraud and 0 otherwise. Logistic Regression is same as linear regression but interpret curve is different. To generalize the linear regression model, when dependent variable is categorical and analyzes relationship between multiple independent variables.

REFERENCES

[1] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum," *2009 International Joint Conference on Artificial Intelligence*, Hainan Island, 2009, pp. 353-356.

[2] Vijayshree B. Nipane, Poonam S. Kalinge, DipaliVidhate, Kunal War, Bhagyashree P. Deshpande, Fraudulent Detection in Credit Card System Using SVM & Decision Tree.

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

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

[5] Snehal Patil, HarshadaSomavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar, Credit Card Fraud Detection Using Decision Tree Induction Algorithm.

[6] E. Aleskerov, B. Freisleben, and B. Rao, " CARDWATCH: A neural network based database mining system for credit card fraud detection," in Proc. IEEE/IAFE Computat. Intell. Financial Eng., Mar. 1997, pp. 220–226.

[7] C. Alippi, G. Boracchi, and M. Roveri, "A just-in-time adaptive classification system based on the intersection of confidence intervals rule," Neural Netw., vol. 24, no. 8, pp. 791–800, 2011.

[8] C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 2, pp. 246–258, Feb. 2016.

[9] C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 4, pp. 620–634, Apr. 2013.

[10] B. Baesens, V. Van Vlasselaer, and W. Verbeke, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Hoboken, NJ, USA: Wiley, 2015.

[11] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," Expert Syst. Appl., vol. 42, no. 19, pp. 6609–6619, 2015.

12] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Detecting credit card fraud using periodic features," in Proc. 14th Int. Conf. Mach. Learn. Appl., Dec. 2015, pp. 208–213.

[13] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Syst., vol. 50, no. 3, pp. 602–613, 2011.

[14] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in Proc. SDM, vol. 7. 2007, pp. 443–448.

[15] R. Bolton and D. Hand, "Statistical fraud detection: A review," Stat. Sci., vol. 17, no. 3, pp. 235–249, 2002.