United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST & CART ALGORITHM

M VENKATA SIVA[1], M PAVANKUMAR[2], M VAMSI[3], MCHAITANYA KRISHNA[4]

[1,2,3,4] UG scholars, Department of AI&ML
DSRK College of Engineering, Vijayawada, India.
[1]sivamuthukuru22@gmail.com, [2]pavanluck2@gmail.com,
[3]vamsimiddela@gmail.com,[4]vamsimiddela@gmail.com,[4]chaitanyakrishnamadipatla777@gmail.com

ABSTRACT: The project is mainly focussed on credit card fraud detection in real world. A phenomenal growth in the number of credit card transactions, has recently led to a considerable rise in fraudulent activities. The purpose is to obtain goods without paying, or to obtain unauthorized funds from an account. Implementation of efficient fraud detection systems has become imperative for all credit card issuing banks to minimize their losses. One of the most crucial challenges in making the business is that neither the card nor the cardholder needs to be present when the purchase is being made. This makes it impossible for the merchant to verify whether the customer making a purchase is the authentic cardholder or not. With the proposed scheme, using random forest algorithm the accuracy of detecting the fraud can be improved can be improved. Classification process of random forest algorithm to analyse data set and user current dataset. Finally In this paper 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.optimize the accuracy of the result data. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision. Then processing of some of the attributes provided identifies the fraud detection and provides the graphical model visualization. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision.

1.1 INTRODUCTION

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 significantly difficult, but also 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 identified as a successful measure for fraud detection. A large amount of data is transferred during online transaction processes, resulting in a binary result: genuine or fraudulent. Within the sample fraudulent datasets, features are constructed. These are data points namely the age and value of the customer account, as well as the origin of the credit card. There are hundreds of features and each contributes, to varying extents, towards the fraud probability. Note, the level in which each feature contributes to the fraud score is generated by the artificial intelligence of the machine which is driven by the training set, but is not determined by a fraud analyst. So, in

regards 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 shrink, the contribution level would parallel. Simply make, these models self-learn without explicit programming such as with manual review. Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms. We use supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction in online or by offline. Random forest is advanced version of Decision tree. Random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by picking only a subsample of the feature space at each split. Essentially, it aims to make the trees de- correlated and prune the trees by fixing a stopping criteria for node splits, which I will be cover in more detail later.

II.EXISTING SYSTEM

In existing System, 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. 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 analgorithm is around 50%. 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.

DISADVANTAGES

1. In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed.
2. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure**.**

III.PROPOSED SCHEME

In proposed System, we are applying random forest algorithm for classification of the credit card dataset. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over 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 selected from a random subset of the full 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 provide a good estimate of the generalization error and to be resistant to over fitting.

3.1 ADVANTAGES OF PROPOSED SYSTEM

Random forest ranks the importance of variables in a regression orclassification problem in a natural way can be doneby 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 case (fraud) and 0 for negative case (not fraud).

IV.LITERATURE SURVEY

Sudha mathy G: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R, vol. 8, no-5, pp. 1954-1966.

Now adays there are many risks related to bank loans, especially for the banks so as to reduce their capital loss. The analysis of risks and assessment of default becomes crucial thereafter. Banks hold huge volumes of customer behaviour related data from which they are unable to arrive at a judgement if an applicant can be defaulter or not. Data Mining is a promising area of data analysis which aims to extract useful knowledge from tremendous amount of complex data sets. In this paper we aim to design a model and prototype the same using a data set available in the UCI repository. The model is a decision tree based classification model that uses the functions available in the R Package. Prior to building the model, the dataset is pre-processed, reduced and made ready to provide efficient predictions. The final model is used for prediction with the test dataset and the experimental results prove the efficiency of the built model.

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.Support Vector Machines (SVM like Neural Networks, is a method based on automatic learning from examples or machine learning, and is becoming very popular in researches as it gives promising opportunities in this field. This method has been developed with the means of adapting it for industrial applications and solutions and soon was applied in many statistical and intelligent fields, such as regression, time series analysis, pattern recognition systems and etc. It has been widely applied in different sciences, such as bioinformatics, text and document classification, pattern recognition, image recognition. One can also find a lot of examples of its application in finance and related sciences. The purpose of this article is to shortly describe the method itself, analyze current researches in credit risk evaluation and bankruptcy prediction and to discuss the further possibilities of its application in this field.

IV.ALGORITHMS

4.1 RANDOM FOREST ALGORITHM

The random Forest algorithm works in several steps:

1. Random Forest builds multiple decision trees using random samples of the data. Each tree is trained on a different subset of the data which makes each tree unique.
2. When creating each tree the algorithm randomly selects a subset of features or variables to split the data rather than using all available features at a time. This adds diversity to the trees.

3. Each decision tree in the forest makes a prediction based on the data it was trained on. When making final prediction random forest combines the results from all the trees.
4. For classification tasks the final prediction is decided by a majority vote. This means that the category predicted by most trees is the final prediction.
5. For regression tasks the final prediction is the average of the predictions from all the trees.
6. The randomness in data samples and feature selection helps to prevent the model from overfitting making the predictions more accurate and reliable.

4.2 CART ALGORITHM

Classification and Regression Trees (CART) is a decision tree algorithm that is used for both classification and regression tasks. It is a supervised learning algorithm that learns from labelled data to predict unseen data.

1. **Tree structure:** CART builds a tree-like structure consisting of nodes and branches. The nodes represent different decision points, and the branches represent the possible outcomes of those decisions. The leaf nodes in the tree contain a predicted class label or value for the target variable.
2. **Splitting criteria:** CART uses a greedy approach to split the data at each node. It evaluates all possible splits and selects the one that best reduces the impurity of the resulting subsets. For classification tasks, CART uses Gini impurity as the splitting criterion. The lower the Gini impurity, the more pure the subset is. For regression tasks, CART uses residual reduction as the splitting criterion. The lower the residual reduction, the better the fit of the model to the data.
3. **Pruning:** To prevent overfitting of the data, pruning is a technique used to remove the nodes that contribute little to the model accuracy. Cost complexity pruning and information gain pruning are two popular pruning techniques. Cost complexity pruning involves calculating the cost of each node and removing nodes that have a negative cost. Information gain pruning involves calculating the information gain of each node and removing nodes that have a low information gain.

How does CART algorithm works?

1. The CART algorithm works via the following process:
2. The best-split point of each input is obtained.
3. Based on the best-split points of each input in Step 1, the new "best" split point is identified.
4. Split the chosen input according to the "best" split point.
5. Continue splitting until a stopping rule is satisfied or no further desirable splitting is available.

United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

V.RESULTS

VI.CONCLUSION

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 requiresmore preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data.

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.

[4] AmlanKundu, SuvasiniPanigrahi, ShamikSural, 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 Comp

[6] Sitarampatel, SunitaGond , "Supervised Machine (SVM) Learning for Credit Card Fraud Detection, International of engineering trends and technology, vol. 8, no. -3, pp. 137- 140, 2014.

[7] SnehalPatil, HarshadaSomavanshi, JyotiGaikwad, AmrutaDeshmane, RinkuBadgujar," Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 92-95

[8] Dahee Choi and Kyungho Lee, "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System", vol. 5, no. - 4, December 2017, pp. 12-24.