Unfolding the Credit Card Fraud Detection Technique by Implementing SVM Algorithm

Bhavesh Gujar[1], Ankush Ginjari[2], Sushant Phase[3], Ashutosh Singh[4], Prof. Chitralekha Dwivedi[5]

[1]*Student, Dept. of Information Technology, Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India*
[2]*Student, Dept. of Information Technology, Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India*
[3]*Student, Dept. of Information Technology, Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India*
[4]*Student, Dept. of Information Technology, Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India*
[5]*Assistant Professor, Dept. of Information Technology, Dr. D.Y. Patil Institute of Technology, Pune, Maharashtra, India*

***

Abstract - Credit card fraud is the most prevalent problem in today's society, and it must be addressed immediately. "The act of washing filthy money such that the source of cash can no longer be identified is known as credit card fraud." Credit card fraud is difficult to detect since massive financial transactions involving large sums of money occur on a regular basis in the global market. The (Anti-credit card fraud Suite) was introduced to detect suspicious activity, although it only applies to individual transactions, not bank account transactions, as previously claimed. We describe a machine learning method based on 'Structural Similarity,' which finds common qualities and behaviours among bank account transactions to address these issues. Due to the difficulties of detecting credit card fraud transactions from large datasets, we provide case reduction algorithms that reduce the input dataset before locating pairs of transactions with other bank accounts that have comparable attributes and behaviours.

Key Words: SVM, Machine Learning, pre-processing, Classification, deep learning.

1. INTRODUCTION

Credit is used to obtain a loan and then repay it over a set period of time. Credit ratings are used to determine how likely a person is to repay their debts, and they must be utilised in order to give money to someone who may be a debtor. [1]. When credit scoring is employed, it is critical that the information regarding debtors be appropriately classified. You must perform a math computation in order to extract relevant information from the data. Data mining is a discipline of science that aims to extract useful data and knowledge from massive datasets. [2]. It is one of the data mining approaches for categorising items into groups. Things are classified in a variety of ways, including Decision Trees and Support Vector Machines. This research will employ the C4.5 algorithm. Quinlan (1996) proposed the C4.5 algorithm as an improved version of the ID3 algorithm. Only categorical (nominal or ordinal) type features can be used to create the decision tree in ID3. However, numerical types can be utilised to construct the tree (intervals or ratios can not be used). This is an example of how the modifications to ID3 in C4.5 enable it to handle numeric types.

Credit receipts are created by categorising debtors into two groups: those who have good credit and those who have negative credit. [4]. In this investigation, credit datasets from Germany were used (GDC). A dataset is a collection of items and their associated properties. The features of an object are the characteristics that distinguish it from others. [5] The German Credit Dataset is available in the UCI Machine Learning Repository. This data is being worked on by Dr. Hans Hofmann, a professor at the University of Hamburg. There are 20 different features, 1000 different occurrences, and two different credit categories in the German Credit Datasets.

The purpose of this work is to increase the C4.5 algorithm's accuracy. The primary purpose of this research is to reduce GCD features. This paper proposed feature distribution and feature splits in datasets. In the preprocessing stage, split feature reduction is used. It's employed since GCD includes 20.000 data points, each with 20 features and 1000 inferences. To extract the best features and improve accuracy, feature reduction is done [6]. Bagging Ensemble [7] is also used to select a suitable ensemble for the C4.5 algorithm. Muslim et al. discovered that by using the Bagging Ensemble, the C4.5 approach can be improved. People will compare the C4.5 algorithm to how it would work on its own to evaluate how this inquiry compares.

2. PROBLEM STATEMENT

The purpose of this project is to identify illegal credit card and bank transaction activity. In the early stages of development, it is still in its youth. To decrease the level of criminal activities. Transmitting funds for purposes other than what they were designed appears to be a common practise in the Credit Card industry.

3. LITERATURE SURVEY

Hongwei Chen1 , He Ai , Zhihui Yang1 , Weiwei Yang1 , Zhiwei Ye1 , Dawei Dong," An Improved XGBoost Model Based on Spark for Credit Card Fraud Prediction.''[1] Many financial organisations suffer significant losses due to credit card theft. Credit card fraud data is so dense that an updated XGBoost model built on Spark is needed to deal with the data imbalance. Balanced training sets were created using the Smote algorithm. The fraud detection system made use of the Spark-based XGBoost classifier. After that, the test sets were sorted out one by one. As part of a model comparison experiment, we compared this project's XGBoost algorithm to other algorithms such as logistic regression and decision trees as well as random forests. There is a 9.1 percent difference between the model provided in this project and the model rated second in terms of Recall, F1-Score and AUC in the trial data. The speedup on the datasets of 70,000, 140,000, and 280,000 samples is 2.06, 3.28, and 3.75, respectively, in the experiment with increased throughput. Experiment findings show that the proposed model accurately and efficiently predicts credit card theft and has a practical effect.

Bing Zhu, Wenchuan Yang, Huaxuan Wang, Yuan Yuan," A Hybrid Deep Learning Model for Consumer Credit Scoring"[2] Consumer credit scoring is an essential part of data mining for risk management of consumer credit, and a variety of data mining methodologies have been created or used to it. Image recognition, computer vision, and other fields have experienced a rise in the use of deep learning techniques in recent years. In this essay, we are using deep learning to improve credit ratings for consumers. The procedure of deciding on features Using a convolutional neural network and relief, we've created a new hybrid model. In studies done on a real-world dataset from a local consumer loan organisation, the proposed model beats other benchmark models like logistic regression and random forest.

Much Aziz Muslim, Aldi Nurzahputra, Budi Prasetiyo" Improving Accuracy of C4.5 Algorithm Using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction"[3] Whether or not a prospective consumer is granted credit is determined by the existence of credit scoring. To accurately classify a debtor, credit scores must be accurate enough. It is possible to classify data in various ways, including the decision tree. Decision tree algorithms such as the C4.5 algorithm can be used. This study aims to improve the C4.5 algorithm's ability to anticipate credit receipts. Accuracy is improved with the Split Feature Reduction Model and the Bagging Ensemble. It is used in the pre-processing step to divide datasets in the number of parts n. Four categories of data were used to create this article. There are 16 features in Split 1; 12 features in Split 2; 8 features in Split 3; and 4 features in Split 4. The C4.5 algorithm is then applied to each split. The C4.5 algorithm and the split feature reduction model yielded an accuracy rate of 73.1% in Split 3. 75.1 percent in Split 3 is the best accuracy achieved by employing the split feature reduction model and bagging ensemble with the C4.5 algorithm. The combined use of a split feature reduction model and a bagging ensemble improved accuracy by 4.6% over the C4.5 algorithm alone.

MILLER ARIZA1,2, JAVIER ARROYO1,3, ANTONIO CAPARRINI4, and MARÍA-JESÚS SEGOVIA." Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending"[4] Peer-to-peer lending necessitates credit risk models that are both efficient and transparent. Traditional machine learning algorithms perform well in terms of prediction but fall short in terms of explanatory capability in the vast majority of cases. However, recent explainability methods, such as the SHAP values, can be employed to circumvent this problem. In this work, we test the well-known logistic regression model against a variety of machine learning algorithms for assigning score in peer-to-peer lending. According to the comparison, the machine learning option outperforms in terms of classification performance as well as explainability. SHAP values, in particular, demonstrate how machine learning can account for dispersion, nonlinearity, and structural fractures in feature-target variable interactions. Our findings indicate that machine learning can be given credit scoring models be both accurate and transparent.

Mary Frances Zeager, Aksheetha Sridhar, Nathan Fogal, Stephen Adams, Donald E. Brown, and Peter A. Beling," Adversarial Learning in Credit Card Fraud Detection"[5] Credit card theft is a severe problem for many financial institutions, costing them billions of dollars each year. Because fraud detection methods do not incorporate information on the adversary's understanding of the fraud detection mechanism, many adversaries continue to avoid detection. The purpose of this research is to develop a dynamic fraud detection system that incorporates information about the "fraudster's" objectives and knowledge base. We employ a game theoretical adversarial learning technique in this research to model a fraudster's best strategy and adjust the fraud detection system ahead of time to better identify future fraud cases. Using a logistic regression classifier as the fraud detection method, we first evaluate the enemy's optimal approach based on the amount of undetected fraud cases.

4. PROPOSED SYSTEM

This section goes into great detail about the proposed ML framework. The framework is constructed in such a way that, given a collection of bank accounts and transactions, it will provide a list of probable money laundering account groups. The framework begins by examining the input and looking for transactions that match. Matching transactions are financial transactions that share characteristics such as deposit and withdrawal amounts. Following that, the framework constructs a network representation of all matching transactions. It then use network-based algorithms to eliminate redundant accounts and transactions. The approach then uses a clustering algorithm to identify questionable ML communities in the network. The term "proposed system" refers to all components, including hardware and software, included in the Respondent's proposal. The Proposed System should include Document Management, Workflow, and Customer Relationship

Management framework's efficiency. Then, using structural similarities, we can identify and group possible credit card accounts. Our preliminary experimental results show that we can detect ML accounts with a high degree of accuracy.The prior information on the dataset, such as its attributes, dimensions, and data types of each feature, etc., is an essential factor that helps one to perform proper operations. An offline dataset which is a publically accessible web platform named "Kaggle" is considered for the implementation of the program. The dataset is a Credit card fraud dataset that consists of several transactions. The dataset contains a combination of cases of fraud and non-fraud. CSV files are the most commonly used format for machine learning data. The dataset contains rows and columns of the following features like Merchant_id, Transaction amount, Is declined, Total Number of declines per day, is Foreign Transaction, is HighRisk Country, and is Fraudulent.

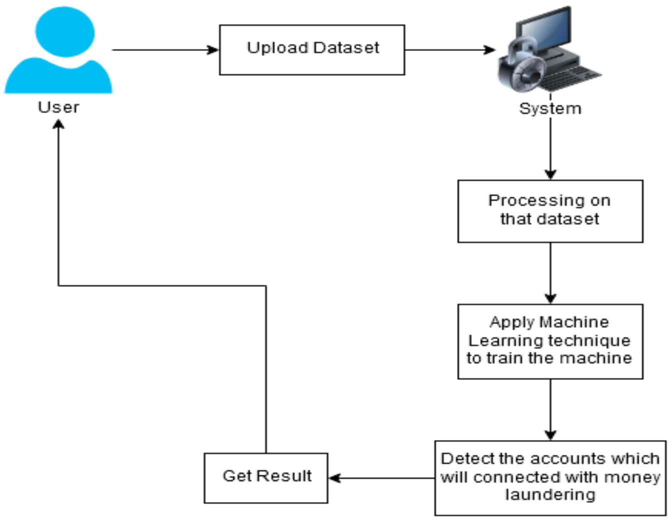
User
Upload Dataset
System
Processing on that dataset
Apply Machine Learning technique to train the machine
Detect the accounts which will connected with money laundering
Get Result

Figure 1. System Architecture

5. ALGORITHM

SVM (Support Vector Machine)

The hyperplane (basically a two-dimensional line) that involves different the tags is created by a support vector machine using these data points. This connect as a decision boundary: anything on one side is categorized as blue, while everything on the other is classified as red. These data points are used by a support vector machine to build a hyperplane (basically a two-dimensional line) with various tags. This link serves as a decision boundary: anything on one side is labelled blue, while everything on the other side is labelled red. SVMs are supervised machine learning methods that can be used to tackle classification and regression problems. It alters your data using a technique known as the kernel trick before determining a suitable border between the available outputs based on these adjustments. The goal of the SVM algorithm is to find a hyper - plane in an N-dimensional space that categorises the input points clearly. The number of features determines the size of the hyperplane. The hyperplane is essentially a line if there are only two input features. The goal of the SVM algorithm is used to find a hyperplane in an N-dimensional space that categorises the input points clearly. The number of features determines the size of the hyperplane. The hyperplane is essentially a line if there are only two input features. SVM can be employed when the number of features in the dataset is large in contrast to the number of data points. By using the appropriate kernel and configuring the best set of parameters. SVM is an excellent, but not the best, classifier. In truth, no one can claim to be the best. SVM works reasonably effectively when there is a clear margin of distinction.

6. CONCLUSION

The proposed machine learning system aims to identify possible Credit card fraud groups within a huge number of economic transactions. Case reducing methods such as matching transaction detection and balance score filter are used to restrict the list of prospective ML accounts in order to improve the framework's efficiency. Then, using structural similarities, we can identify and group possible credit card accounts. Our preliminary experimental results show that we can detect ML accounts with a high degree of accuracy.

REFERENCES

[1] Rokach, Lior and Maimon, Oded. Data Mining with Decision Trees: Theory and Applications 2nd Eddition. Singapore: World Scientific Publishing Co. 2015.

[2] Sugiharti, E. and Muslim, M.A.,. "On-Line Clustering of Lecturers Performance of Computer Science Department of Semarang State University Using K-Meansalgorithm". Journal of Theoretical and Applied Information Technology, 83(1), p.64. 2016

[3] Prasetyo, Eko. Data mining: Mengolah Data menjadi Informasi Menggunakan MATLAB. Yogyakarta: Andi Offsett. 2014.

[4] Nurzahputra, A. and Muslim, M.A., "Peningkatan Akurasi Pada Algoritma C4. 5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit". Prosiding SNATIF, pp.243-247. 2017.

[5] Hermawati, F. A. DATA MINING. Yogyakarta: Andi Offset. 2013.

[6] Wijaya, K. P. & Muslim, M. A. "Peningkatan Akurasi Pada Algoritma Support Vector Machine Dengan Penerapan Information Gain Untuk Mendiagnosa Chronic Kidney Disease". Prosiding 3rd Seminar Nasional Ilmu Komputer. Semarang: Universitas Negeri Semarang. 2016.

[7] Lessmann, S., Baesens, B., Seow, H.V., and Thomas, L.C., 2015. Benchmarking state-of-the-art classification

algorithms for credit scoring: An update of research. European Journal of Operational Research 247, 1, 124-136.

[8] Crook, J.N., Edelman, D.B., and Thomas, L.C., 2007. Recent developments in consumer credit risk assessment. European Journal of Operational Research 183, 3, 1447-1465.

[9] Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2017, pp. 0630-0633.

[10] D. Prusti and S. K. Rath, "Web service based credit card fraud detection by applying machine learning techniques," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 492-497.

[11]Rokach, Lior and Maimon, Oded. Data Mining with Decision Trees: Theory and Applications 2nd Eddition. Singapore: World Scientific Publishing Co. 2015.

[12] Sugiharti, E. and Muslim, M.A.,. "On-Line Clustering of Lecturers Performance of Computer Science Department of Semarang State University Using K-Meansalgorithm". Journal of Theoretical and Applied Information Technology, 83(1), p.64. 2016.

[13] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," Risks, vol. 7, no. 1, 2019.

[14] C. Serrano-Cinca and B. Gutierrez-Nieto, "The use of profifit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," Decision Support Systems, vol. 89, pp. 113–122, 2016.