The Use of Machine Learning for the Determination of Credit Card Transaction Fraud

Dr. G Ahmed Zeeshan[1], Mrs. G Pavani[2], Mrs. Kshetravathi N S[3]
Associate Professor[1], Assistant Professor[2,3]
Department of CSE
Global Institute of Engineering and Technology

Abstract:

Identifying and preventing actual instances of credit card fraud is the primary goal of this research. There has been a dramatic increase in fraudulent activity due to the skyrocketing development of credit card transactions. In today's world, internet use has skyrocketed and is becoming an integral component of daily life. With the growth of e-commerce, purchasing and selling goods online has become more convenient and open to more creative approaches. With the advent of contemporary technologies such as online banking and credit card payments, the use of online bill payment and online shopping has grown substantially in recent years. As the use of credit cards grows, the difficulty for banks to differentiate between legitimate purchases and fraudulent ones has grown in tandem with the popularity of online payment and shopping. Another scenario where credit card theft might occur is if the client misplaces their card. Consequently, banks find it more difficult to halt fraudulent transactions after they have occurred. Using real-time machine learning, this study demonstrates how to identify fraudulent transactions and prevent their processing.

I. Introduction

Researchers have been considering ways to build models based on AI, data mining, fuzzy logic, and machine learning in order to identify fraudulent behaviors in credit card transactions. Detecting credit card fraud is a common but very challenging topic. Our suggested approach incorporates a machine learning-based credit card fraud detector. Thanks to the development of better machine learning methods. When it comes to detecting fraud, machine learning has shown to be an effective approach. Online transactions involve the transmission of vast amounts of data, which can only have two possible outcomes: legitimate or fraudulent. Features are built inside the sample datasets that are false. Here we have certain data points, like the customer's account age and value, along with the credit card's origin. Each of the hundreds of factors affects the likelihood of fraud in its own unique way. Keep in mind that a fraud analyst does not decide how much weight each feature has in calculating the fraud score; rather, this decision is made by the machine learning algorithm that is trained on the training set. Accordingly, with respect to card fraud, the fraud weighting of a credit card transaction will be similarly high if the usage of cards to commit fraud is demonstrated to be significant. But if this were to decrease, the amount of the contribution would be the same. Just like manual reviews, these models can self-learn without any programming at all. Machine learning-based credit card fraud detection makes use of regression and classification techniques. When it comes to online or offline card transactions, we use supervised learning algorithms like the Random Forest algorithm to categorize them. Decision trees have an upgraded form called random forests. When compared to other machine learning methods, random forest is the most efficient and accurate. With each split, random forest selects a smaller subset of the feature space in an effort to mitigate the correlation problem. I will go into more depth later on, but basically, it tries to de-correlate the trees and prune them by establishing a stopping condition for node splits.

II.LITERATURE SURVEY

A Study on Credit Card Fraud Detection Using Predictive Analytics

The purpose of this study is to compile a scorecard for the present state of credit card fraud detection

using predictive analytics vendor solutions by collecting and organizing pertinent assessment criteria, features, and capabilities. This scorecard compares five different vendor solutions for credit card predictive analytics that have been implemented in Canada. A comprehensive inventory of the difficulties, dangers, and restrictions associated with credit card fraud PAT vendor solutions was compiled from the subsequent study results.

"BLAST-SSAHA Hybridization for Credit Card Fraud Detection"

In the first step of the two-stage sequence alignment described in this work, a profile analyzer (PA) checks the authenticity of a credit card's incoming transaction sequence against the actual cardholder's spending sequence. After the profile analyzer identifies suspicious transactions, they are sent to a deviation analyst (DA) to see whether they match any fraudulent patterns from the past. The findings from these two analysts provide the foundation for the ultimate determination about the kind of a transaction. Our proposed method blends the sequence alignment methods BLAST and SSAHA in a novel way, allowing for online response time for both PA and DA.

"Research on Credit Card Fraud Detection Model Based on Distance Sum"

A dramatic increase in credit card fraud has occurred in China in tandem with the country's expanding credit card market and transaction volume. The primary concern of banks' risk control departments is how to better identify and stop credit card fraud. It suggests a technique for detecting credit card fraud that applies outlier mining to credit card transaction data, with outlier detection based on distance sum taking into account the irregularity and rarity of fraudulent transactions. The model is both practicable and accurate, according to the experiments.

"Fraudulent Detection in Credit Card System Using SVM & Decision Tree."

Fraud is becoming more common as e-commerce develops, resulting in huge monetary losses for victims all around the globe. Presently, credit card theft is a big source of financial losses; it impacts both individual consumers and tradespeople. The approaches that are discussed here may be used to identify credit card fraud. They include decision trees, genetic algorithms, neural networks, meta learning strategies, and HMMs. Contemplation uses the artificial intelligence concepts of a Support Vector Machine (SVM) and a decision tree to address the issue of fraudulent identification. Financial losses may be mitigated to a larger extent by using this hybrid technique.

III.SYSTEM ANALYSIS

3.1. EXISTING SYSTEM

In existing System, 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 an algorithm 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.

Drawbacks Of Existing System

- 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.

3.2. PROPOSED SYSTEM

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.

ADVANTAGES OF PROPOSED SYSTEM

- Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by 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.SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture of the proposed technique.

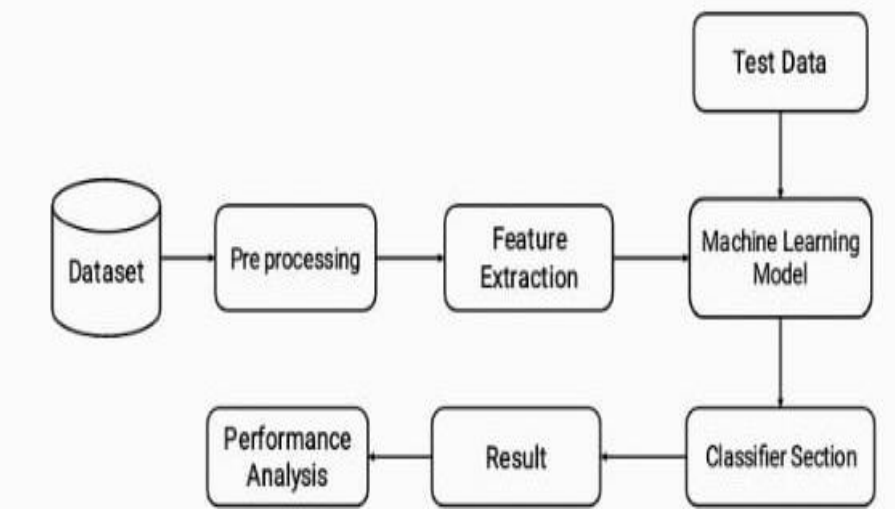

Fig. 4.1.1 System Architecture

V.SYSTEM IMPLEMENTATION

5.1. MODULES

- DATA COLLECTION

- DATA PRE-PROCESSING
- FEATURE EXTRACTION
- EVALUATION MODEL

Module description:

5.1.1 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.

5.1.2 Data Pre-processing

Organize your selected data by formatting, cleaning, and sampling from it. Three common data pre-processing steps are:

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 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.

5.1.3 Feature Extraction

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.

5.1.4 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.

. VI. RESULTS

Fig.6.1 Random Forest accuracy

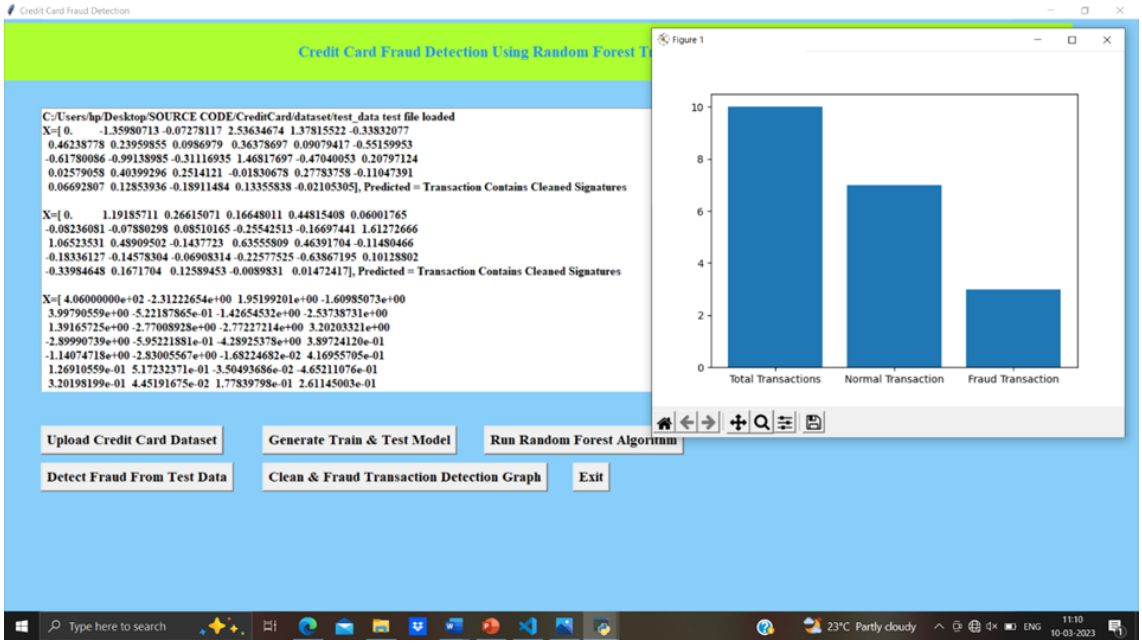

Fig.6.2 Fraud Transaction Detection Graph

VII. CONCLUSION AND FUTURE WORK

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 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. In future we plan to enhance the existing algorithm and train it with another credit card dataset having a greater number of features.

REFERENCES :

Credit Risk Analysis and Prediction Modeling of Bank Loans Using R, vol. 8, no. 5, pp. 1954–1966, as cited in Sudhamathy G. [1].

Credit Risk Assessment for Rural Credit Cooperatives Using an Improved Neural Network, LI Changjian and HU Peng, 2017 International Conference on Smart Grid and Electrical Automation, 60(3), 227–230.

[3] Credit Risk Assessment in Commercial Banks Based on Support Vector Machines, edited by Wei Sun, Chen-Guang Yang, and Jian-Xun Qi, published in 2006, volume 6, pages 2430–2433.

[4] "BLAST-SSAHA Hybridization for Credit Card Fraud Detection," vol. 6, no. 4, pp. 309-315, 2009, AmlanKundu, Suvasini, Panigrahi, Shamik, and Surya, Senior Member, IEEE.

[5] In the 2011 Proceedings of the International Multi Conference of Engineers and Computer Scientists, volume I, Y. Sahin and E. Duman discuss "Detecting Credit Card Fraud using Decision Trees and Support Vector Machines".

[6] "Supervised Machine (SVM) Learning for Credit Card Fraud Detection," The International Journal of Engineering Trends and Technology, volume 8, issue 3, pages 137–140, 2014, by SunitaGond Sitarampatel.

In an article published in the International Journal of Computer Science and Mobile Computing in April 2015, the authors Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, and Rinku Badgujar discuss a method for detecting credit card fraud using a decision tree induction algorithm. The article spans pages 92 to 95.

[8] "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System," Dahee Choi and Kyungho Lee, publication date: December 2017, volume 5, issue 4, pages 12–24.