Detecting of Anomaly Activities in Credit Card Transaction using Machine Learning Algorithm

K. Meena[1], R. Gayathri[2]
[1]Assistant Professor, [2] M.Sc. Student
[1] Department of Software Engineering, [2] Department of Software Engineering
Periyar Maniammai Institute of Science and Technology,
Thanjavur, Tamilnadu, India.

Abstract: The Credit card usage is the one of the important part in today's economy. With the wide use of credit cards the fraud will appears as a major issue in the credit card business. A large number of fraud transactions are made every day. So many techniques are availed for detecting fraud transactions in credit card. Each and every techniques implies their merits, demerits and principles. Today machine learning plays a major in major activities of Artificial Intelligence. Based on machine-learning techniques detecting credit card fraud is to reduce major issues in this transaction. Under the machine learning approach using Decision tree is used to detect credit card fraud in this paper. The experimental results depicts Decision tree based approaches provides highest accuracy when compared with other techniques and this proposed approach will be very useful for fraud investigators.

Index Terms – Credit Card, decision tree, fraud detection, machine learning

1. Introduction

At the current state of the world, financial organizations expand the availability of financial facilities by employing of innovative services such as credit cards, ATM, internet and mobile banking services. Besides, along with the rapid advances of e-commerce, the use of credit card has become a convenience and necessary part of financial life. Credit card is a payment card supplied to customers as a system of payment. There are lots of advantages in using credit cards .There is a rapid growth in the number of credit card transactions which has led to a extensive rise in fraudulent activities. In this paper, we will focus on credit card fraud and its detection measures. A credit card fraud occurs when one individual uses other individuals' card for their personal use without the knowledge of its owner. When such kind of cases takes place by fraudsters, it is used until its entire available limit is useless. Using Machine Learning (ML) techniques we can efficiently discover fraudulent patterns and predict transactions that are probably to be fraudulent. Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set. Decision Tree learning algorithm generates decision trees from the training data to solve classification and regression problem. Decision trees are constructed an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules conditional from the data features. Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers.

2. Related Work

There are different machine learning techniques were used to approach the credit card fraud detection problem. In these authors in [1] framed models both isolate and combined machine learning techniques for handwritten digits' recognition. The results showed that the classification accuracy of the combined classifier model outperformed the individual classifier model. A fraud detection model based on the decision trees was developed in [2] and indicated fitting problem in decision tree. In these authors in [3] they present machine learning method to classifying anomalous and normal activities in a computer network with supervised and unsupervised algorithm that have the best efficiency. They used K-means clustering and the Id3 decision tree which improves the system classification performance. In [4] decision trees and support vector machines (SVM) are applied on a dataset obtained from a real world national bank's credit card data warehouses. They found out that decision trees outperform SVM in solving the problem. [5]Analysis of the dataset is performed using different classification techniques that is K-Mean which is based on clustering, K-Nearest neighbor, Support Vector Machine. To classify and normalize that data first analyzed that flat result then preprocessed data is used and for preprocessing statistical normalization has been used. By applying classification algorithm in data without any preprocessing they generates good result but when data get normalized then degrades potential of classification techniques. They have concluded that after evaluation of all classification algorithm K-Nearest neighbor provides better result as compared to both K-means and SVM but it takes more execution time. In these [6] they used supervised learning neural network with preprocessing step for intrusion detection. To generate the sample for the original dataset he used stratified weighted sampling technique and this sample applied on the proposed algorithm. The result showed the proposed system created higher accurateness and low error in identifying whether the records are usual or attack one. In these [7] they used feature selection approach, for the intrusion detection system should be fast and effective an optimal feature subset should be made. Out of all the subset which they certain out of 41 features, the best presentation is given by the subset of 15 features which is almost equal to the performance given by the set of 41 features and time taken to construct the model by the subset of 15 features is less than the time taken by the set of 41 features. The authors in [8] developed two models based on logistic regression and SVM. In [9] Zero day attack is one of the variety of anomalies and it is very important and real time challenge for both network operators and researchers. They proposed an alternative detection technique which is based on combination of feature space and time series to automatically detect anomalies in real time for using machine learning algorithm. They conducted experiments with real time traffic in real world and compared the detection performance and their result show proposed technique do better in task of anomaly detection and has a good possibility for applying in real time system. In [10] ANN based model and decision trees model are compared, and the authors found that the ANN outperforms decision trees. In [11] it was found that logistic regression outperforms SVM.

3. METHODOLOGIES

Detecting Fraud can be done with binary classification task in which any transaction will be predicted and labeled as a fraud or legit. In this paper the anomaly detection can be done by efficient machine algorithm (i.e.) decision tree for detecting malicious users. A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf the decision tree algorithm writing steps are

1.Start with a training data set which is called as S. It should have attributes and classification.
2.Determine the best attribute in the dataset. Split S into subset that contains the possible values for the best attribute.
3.Make decision tree node that contains the best attribute.
4.Recursively generate new decision trees by using the subset of data created from step 3 until a stage is reached where you cannot classify the data further. Represent the class as leaf node.

In decision tree CART is the one of the algorithm. The CART (Classification and Regression Trees) → uses *Gini Index (Classification)* as metric.By use the Gini Index as our cost function used to evaluate splits in the dataset. Our target variable is Binary variable which means it take two values .Gini index working steps

1. Compute the Gini index for data-set
2. for every attribute/feature:
    1.Calculate Gini index for all categorical values
    2.Take average information entropy for the current attribute
    3.Calculate the Gini gain

3. Pick the best Gini gain attribute.
4. Repeat until the desired tree is available.

3.1 Data cleaning

Data cleaning is the process of modifying data to assure that is correct, accurate, and relevant. It is the process of cleaning / standardizing the data to make it ready for analysis. Most of times, there will be discrepancies in the captured data such as incorrect data formats, missing data, errors while capturing the data. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making. While it has been the focus of many researchers for several years, individual problems have been addressed separately. These include missing value imputation, outlier's detection, transformations, integrity constraints violations detection and repair, consistent query answering, deduplication, and many other related problems such as profiling and constraints mining.

3.2 Feature selection

This module is to select best feature from the given dataset. In machine learning and statistics, *feature selection* means retrieving a subset relevant, useful features to use in building an analytical model. Feature selection helps narrow the field of data to the most valuable inputs. Narrowing the field of data helps reduce noise and improve training performance.

3.3 Machine learning model

3.3.1 Evaluate

This module is to measure the model's accuracy. The model provide scores and the **Evaluate Model** module computes a set of industry-standard evaluation metrics.There are three ways to use the Evaluate Model module:

- Generate scores over your training data, and evaluate the model based on these scores
- Generate scores on the model, but compare those scores to scores on a reserved testing set
- Compare scores for two different but related models, using the same set of data

3.3.2 Initialize

These modules provide the machine learning algorithms, which you can customize by setting parameters

3.3.3 Score

By using score new data can be passed and produce the result. You can also use the results of scoring as part of a predictive service..Scoring is widely used in machine learning to mean the process of generating new values, given a model and some new input. The generic term "score" is used, rather than "prediction," because the scoring process can generate so many different types of values:

- A list of recommended items and a similarity score.
- Numeric values, for time series models and regression models.
- A probability value, indicating the likelihood that a new input belongs to some existing category.
- The name of a category or cluster to which a new item is most similar.
- A predicted class or outcome, for classification models.

3.3.4 Train

This is the first task in machine learning model on data you provide. When Machine Learning is training a model, rows with missing values are skipped. Therefore, if you want to fix the values manually, use imputation, or specify a different method for handling missing values, use the Clean Missing Data module before training on the dataset. We recommend that you use the Edit Metadata module to fix any other issues with the data. You might need to mark the label column, change data types, or correct column names. For other common data cleanup tasks, such as normalization, sampling, binning, and scaling, see the Data Transformation category.

4. Experiments and Results

The fraud detection models were trained and tested using python. Python is an interpreter, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. Python has gained huge popularity and contains special libraries for ML. The language is great to use when working with ML Algorithms and has easy syntax.

The input with the dataset size as 2,50,000 as given. We have used precision, recall, f1-score in the different models. The average performance results are stored. This methodological approach ensures that all data were represented once as a test data and several times as a training data producing accurate results.

Table 1 represents the Performance Metrics of Logistic regression algorithm. Table 2 represents the Performance Metrics of Decision Tree algorithm. Table 3 represents the Performance Metrics of KNN classifier algorithm. Table 4 represents the Performance Metrics of Stochastic Gradient Descent Classifier Table 5 represents Performance Metrics of Passive Aggressive Classifier Table 6 represents the Performance Metrics of Perceptron.

Table 1: Performance Metrics of Logistic regression

	Precision	Recall	F1-Score	Support
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.56	0.23	0.23	103
micro avg	1.00	1.00	1.00	56962
macro avg	0.78	0.62	0.66	56962
weighted avg	1.00	1.00	1.00	56962

Table.2: Performance Metrics of Decision Tree

	Precision	Recall	F1-Score	Support
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.52	0.49	0.50	103
micro avg	1.00	1.00	1.00	56962
macro avg	0.76	0.74	0.75	56962
weighted avg	1.00	1.00	1.00	56962

Table.3: Performance Metrics of KNN classifier

	Precision	Recall	F1-Score	Support
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.25	0.01	0.02	103
micro avg	1.00	1.00	1.00	56962
macro avg	0.62	0.50	0.51	56962
weighted avg	1.00	1.00	1.00	56962

Table.4: Performance Metrics of Stochastic Gradient Descent Classifier

	Precision	Recall	F1-Score	Support
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.00	0.00	0.00	103
micro avg	1.00	1.00	1.00	56962
macro avg	0.50	0.50	0.50	56962
weighted avg	1.00	1.00	1.00	56962

Table.5: Performance Metrics of Passive Aggressive Classifier

	Precision	Recall	F1-Score	Support
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.00	0.00	0.00	103

micro avg	1.00	1.00	1.00	56962
macro avg	0.50	0.50	0.50	56962
weighted avg	1.00	1.00	1.00	56962

Table.6: Performance Metrics of Perceptron

	Precision	**Recall**	**F1-Score**	**Support**
Non Fraud	1.00	1.00	1.00	56859
Fraud	0.00	0.00	0.00	103
micro avg	1.00	1.00	1.00	56962
macro avg	0.50	0.50	0.50	56962
weighted avg	1.00	1.00	1.00	56962

The figure. 1 shows the output of credit card transaction of sub dataset size 10000 as Decision Tree. This tree shows the fraud and non fraud transaction as clearly.

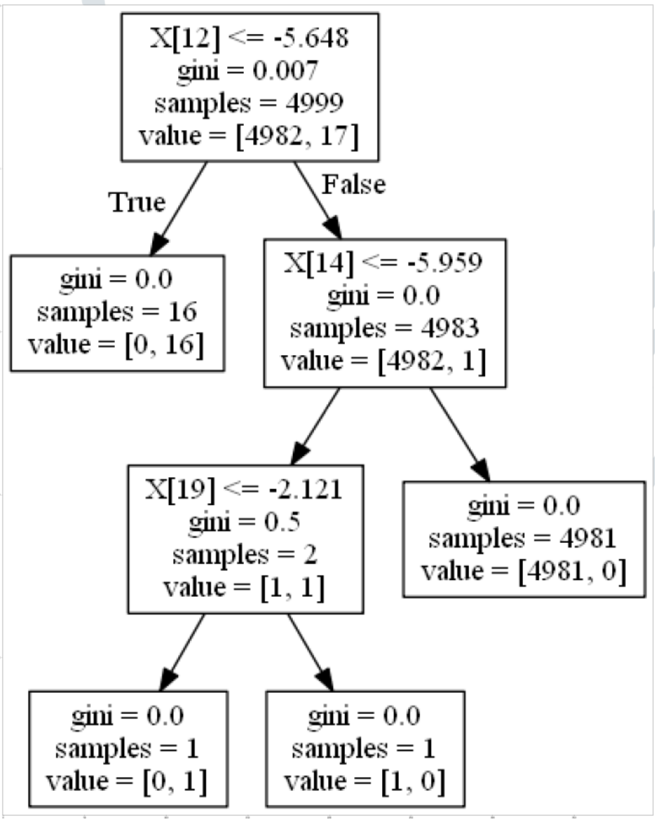

Figure. 1: Decision Tree for credit card transaction of sub dataset size 10000

5. Conclusion

In this paper Decision tree, based machine learning approach is used to detect credit card fraud. The results show that there performance of Decision tree based approaches beats with the highest accuracy and it can be effectively used for fraud investigators. In future with different machine learning techniques the performance can be enhanced for credit card fraud investigation.

References

[1] Ng, G., & Singh, H. (1997). Democracy in pattern classifications: combinations of votes from various pattern classifiers. Nan yang, Singapore: Elsevier

[2] Ehramikar, S. (2000). The Enhancement of Credit Card Fraud Detection Systems. Toronto, Canada: Master of Applied Science The-sis, University of Toronto. V. Vapnik. Statistical Learning TheoryWiley, New York, 1998.

[3] K. Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikas Krishna, “Implementation of Anomaly Detection Technique using Machine Learning Algorithms” international journal of computer science and telecommunications 2011.

[4] Sahin, Y., & Duman, E. (2011). Detecting Credit Card Fraud by De-cision Trees and Support Vector Machines. Hong Kong, China: TheInternational MultiConference of Engineers and Computer Scien-tists.

[5] Riti Lath, Manish Shrivastava, “Analytical Study of Different Classification Technique for KDD Cup data’99” international journal of applied information system 2012.

[6] Devendra kailashiya, Dr. R.C. Jain, “Improve Intrusion Detection using Decision tree with Sampling” international journal computer technology and applications 2012.

[7] Sneh Lata Pundir and amrita, “Feature selection using Random forest in intrusion detection system” international journal of advances in engineering and technology 2013.

[8] Huang, S. (2013). Fraud Detection Model by Using Support Vector Machine Techniques. Chiayi, Taiwan: International Journal of Digi-tal Content Technology & its Applications

[9] Kriangkrai limthong, “Real Time Computer Network Anomaly Detection using Machine Learning Techniques” journal of advances in computer networks 2013.

[10]Zaki, M., & Meira, W. (2014). Data Mining and Analysis: Funda-mental Concepts and Algorithms. New York City, New York: Cam-bridge University Press.https://doi.org/10.1017/CBO9780511810114.
.
[11]Balaji, G. N., T. S. Subashini, and N. Chidambaram. "Detection ofheart muscle damage from automated analysis of echocardiogramvideo."IETE Journal of Research61.3 (2015): pp: 236-243.https://doi.org/10.1080/03772063.2015.1009403.