Machine Learning for Credit Card Fraud Detection

[1]Sameh Gamal Khalil Taktak, [2]Atef Zaki Ghalwash, [3]Amr Galal and [4]Mohamed M. Abbaassy
[1]Department of Business Information System, Faculty of Commerce and Business Administration, Helwan University ,Cairo, Egypt
[2]Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt
[3]Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt
[4]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Cairo, Egypt

Key words: Credit fraud detection, machine learning, SVM, decision tree , logistics regression

Corresponding Author:
Atef Zaki Ghalwash
Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

Page No.: 06-10
Volume: 21, Issue 2, 2022
ISSN: 1682-3915
Asian Journal of Information Technology
Copy Right: Medwell Publications

Abstract: Due to the increased use of credit cards for online purchases, fraud has increased exponentially because of the rapid rise in the e-commerce business. In recent years, banks have found it increasingly difficult to detect credit card fraud. The power of machine intelligence can detect credit card fraud. Banks have used a variety of machine learning approaches, prior data and novel attributes to better forecast these transactions. For credit card transactions, the sampling method used, as well as the selection of data points and the detection techniques used, can all have a significant impact on fraud detection. Credit card fraud is investigated using logistic regression, decision trees, random forests and Support Vector Machines (SVM). In September 2013, the data included all transactions made by European cardholders. There were 492 instances of fraud out of a total of 284, 807 transactions. It classifies fraudulent transactions as "positive" and legitimate transactions as "negative.". Fraud accounts for 0.173% of the total transactions in the data set, making it highly imbalanced. To balance the data set, over sampling was used, resulting in 60% of fraudulent transactions and 40% of genuine transactions. Among the four models, Logistic Regression produced the best results, the Logistic Regression model has a 96% accuracy.

INTRODUCTION

Credit card fraud is a broad term that refers to theft and fraud committed with or involving a credit card at the time of payment. The goal could be to purchase goods without paying or to withdraw unauthorized funds from an account. Credit card fraud is a complication of identity theft. According to the United States Federal Trade Commission, the rate of identity theft remained stable during the mid-2000s, but it increased by 21% in 2008. Even though credit card fraud, the crime most people associate with ID theft, has decreased as a percentage of all ID theft complaints, In the year 2000, approximately 10 million, or one out of every 1300 transactions, were found to be fraudulent out of 13 billion transactions made annually. One-twelfth of one percent of all transactions are now monitored by fraud detection systems, resulting in billions of dollars in losses. Credit card fraud is now one of the most serious threats to business enterprises. However, to effectively prevent fraud, it is necessary to first understand how fraud is carried out. Credit card scammers use a variety of techniques to defraud people. Credit card fraud is defined as "when an individual uses another individual's credit card for personal reasons while the card owner and card issuer are unaware that the card is being used." The first step in card fraud is the theft of the actual card or critical data associated with the account, such as the card account number or other information that must be available to a merchant during a valid transaction.

As technology advances, there is an increase in the number of frauds. Logistic regression, decision trees, random forests and support vector machines are increasingly being used to detect fraud due to their superior performance.

Credit card fraud is on the rise and as a result, financial losses are rapidly rising. Each year, billions of dollars are lost because of scams. There is a scarcity of data to properly investigate the fraud. Detecting real-world credit card fraud necessitates the application of a variety of machine learning techniques. Logistic regression, decision trees, random forests and support vector machines are among the algorithms employed.

Related work: Machine learning is widely used in a variety of high-efficiency data processing fields, including the detection of card fraud, to name a few. As a result of the need to understand some of the technologies involved in credit card fraud identification, as well as a better understanding of the various types of card fraud, several approaches to detecting fraud have been proposed. These approaches range from unsupervised detection strategies to a hybrid approach. Several different approaches were tried and tested[1]. Training and learning from transaction data are required to identify new frauds and this represents a significant portion of the process for detecting and preventing fraud[2-3]. As a result, it is critical to create a model that incorporates the best Data Mining and Machine Learning algorithms available to detect fraud quickly and take immediate preventive measures. A well-designed model would not only be able to identify frauds with high reliability, but it would also be able to predict the possibility of future fraudulent behavior, which would be extremely beneficial.

When a cardholder completes a transaction, the cardholder's actions are examined for evidence of fraud[4]. Most strategies, such as artificial neural network (ANN), genetic algorithm (GA), support vector machine (SVM), frequent item set mining (FISM), decision tree (DT), optimization algorithm for migratory birds (MBO) and process for naive Baiyes, were used in the identification of card fraud (NB). The quantitative logistic regression and naive bays analysis are carried out in. On credit card fraud data, the output of Bayesian and neural systems is evaluated[5].

Proposed model: In this study, the algorithms given are used to detect credit card fraud. There are several machine learning algorithms that may be used by credit card merchants to identify fraudulent transactions, including Logistic Regression, Decision Trees and Random Forest. The complete system architecture is shown as in the figure shown in Fig. 1.

Logistic regression: When the dependent variable is binary, it is an appropriate technique for predictive analysis. This technique has the potential because classifying transactions as fraud has this double edge. Using a logistic curve, this statistical classification model detects fraud. This logistic curve can be used to evaluate class membership probability because its value ranges from 0 to 1.

70% of the predictions were correct. Since a single line can split the plane using threshold probabilities and divide the dataset points into exactly two parts, logistic regression is a good choice. As a result, it is difficult to deal with outliers. To determine the probability of an event, it makes use of a natural logarithmic function.

Decision tree algorithm: A new algorithm for supervised learning there is many ways to split trees, but the most common method is to use the binary or multi-split method to create child nodes. Each tree uses its own algorithm to perform the splitting process until there is no more splitting needed in our model[6]. Each attribute is assigned a value based on the input variables associated with the method being used and each node is associated with an input variable relevant to the method being used.

Overfitting of the training data may occur as the tree grows, with possible anomalies in branches, errors, or noise. As a result, pruning is used to improve a tree's classification performance by removing certain nodes from its structure. Decision trees are widely used because they are simple to use and can handle a wide range of data.

Random forest: One of the reasons for the development of random forests was the instability and sensitivity of single trees to some training data. Random forests are more efficient since each tree is constructed independently of the others[7].

In principle, it is an ensemble of regression and/or classification trees, which are straightforward to use because they only use two randomness sources, or parameters, that are bootstrapped along samples and consider only a random data attribute subset to build each tree.

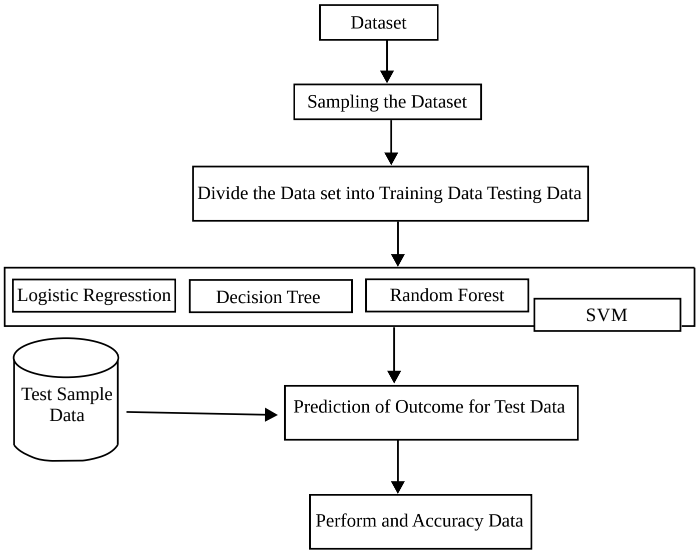
Dataset
Sampling the Dataset
Divide the Data set into Training Data Testing Data
Logistic Regresstion
Decision Tree
Random Forest
SVM
Test Sample Data
Prediction of Outcome for Test Data
Perform and Accuracy Data

Fig. 1: Proposed system architecture

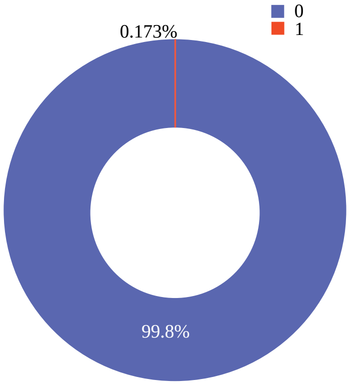

Fig. 2: Check balance and analysis the target

SVM: Since a non-linear task in the input gets transformed into a linear one when dimensionality increases, Support Vector Machines (SVMs) are linear classifiers that can be used to detect fraud[8].

It has a high level of adaptation because of its two most essential features: a kernel function for representing classification functions in the dot product of input data point projection and a hyperplane to maximize separation between classes while limiting overfitting of training data.

lsImplementation: The dataset represents all the transactions made by European cardholders in September 2013. The 284,807 total transactions, 492 were fraud. In comparison to the huge amounts of transaction data, there are fewer occurrences of fraud because that data is not balanced. There are no text values in this dataset, as it was transformed using PCA. Due to concerns about privacy and confidentiality, only PCA-transformed data is supplied. Other than time and money, all the given values v1, v2, v3 and v28 are PCA-transformed numbers. When a transaction is flagged as fraud, the feature class value is 1 and when it is not, the value is 0.

Understanding the problem statement and data, performing statistical analysis and visualization and determining if the data is balanced or not are the steps, we used to predict the result as shown in Fig. 2. Using oversampling with data that is imbalanced, standardization and normalization, the data in this dataset is balanced before being tested with a variety of machine learning techniques.

Training and testing datasets are separated in the dataset. Seventy percent of the data is being used for training, while the other 30% is being used for test purposes. The algorithms are SVM, Logistic Regression, Decision Tree and Random Forest with a boosting approach.

After the dataset has been trained on, the testing procedure begins. The results of the tests on each algorithm's performance will be represented visually in the form of graphs. The optimal algorithm is selected based on the accuracy of the outputs from each algorithm.

In terms of evaluation, each algorithm has its own set of performance measures and these measures have been designed to evaluate a wide range of items. As a result, different ways are being reviewed and this should be one of the criteria considered. It is common in credit card fraud detection to use the False Positive (FP) and False Positive (FP) as well as their relationship to compare the accuracy of different approaches. The following sections will provide definitions for the parameters that we previously discussed.

A true positive: In the situation where a transaction is accurately identified as fraudulent, it is referred to as a "true positive" (TP):

$$\text{True positive} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

True negative: The true negative rate is the proportion of normal transactions that are accurately identified as normal transactions, which is calculated as follows:

$$\text{True negative} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

False positive (FP): The proportion of non-fraudulent transactions that are incorrectly categorized as fraudulent transactions, which is calculated as follows:

$$\text{False positive} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

False negative (FN): It means a proportion of non-fraudulent transactions are being misclassified as normal transactions:

$$\text{False negative} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

More insight into the predictive model's performance, as well as which classes have been correctly predicted and which have been incorrectly predicted, can be gained by using a confusion matrix. Two-class classification problem with negative and positive classes is the simplest confusion matrix. Each cell in the table has a well-known name in this type of confusion matrix.

Accuracy is defined as the percentage of instances that are correctly classified. One of the most used metrics for classifying performance in a classification system:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}}$$

To use binary classification models, for instance, The definition of accuracy is:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Precision: The number of classified positive or fraudulent instances that are positive instances:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: Recall is a metric that measures the proportion of correctly predicted positive outcomes among all possible outcomes. In contrast to precision, which only comments on positive predictions that are correct, recall provides an indication of those that are incorrect. False negatives are counted as one if there are more of them than there are positives and vice versa:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score: F1 score is calculated as the weighted average of precision and recall. As a result, this score records both false positives and false negatives:

$$\text{F1 Score} = 2\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Support: The number of occurrences of a class in a dataset. There are a certain number of samples of the true response in that class. A classifier's reported scores may be affected if the training data is unbalanced or if stratified sampling or rebalancing is required. No matter which model is being evaluated, support remains constant.

RESULTS AND DISCUSSIONS

Models such as logistic regression, SVM, decision tree and random forest with boosting technique performed well against the data, as evidenced by the following findings.

Grouping this dataset is a challenge because fraudulent and lawful transactions are so similar. As a result, it is highly difficult to separate the fraud cases from the genuine groups.

The comparison Table 1 was created using the computer simulation results. The factors that are compared are accuracy, precision and recall. The table shows that the Logistic Regression model has the highest accuracy, precision and recall.

The Logistic Regression gave the best result among the four models. The accuracy rate for the Logistic Regression model is 96 %. The Decision Tree model performs the worst. Because the dataset we've gathered is classified. The dependent attribute named class should be used to divide the dataset into 0s and 1s.

Table 1: Accuracy, precision and recall of various machine learning algorithms

Parameters	Accuracy (%)	Precision (%)	Recall
SVM	93	88	98%
Random forest	95	93	1
Logistic regression	96	95	1
Decision tree	92	92	91%

CONCLUSION

Machine learning techniques such as logistic regression, random forests, decision trees and SVM were used to detect credit card fraud. The performance of the proposed system is measured using several metrics, including sensitivity, specificity, accuracy and the rate at which errors occur. Each of the four models has an overall accuracy rate of 96, 92, 95 and 93%. When compared to the other four models, logistic regression outperforms random forests, decision trees and support vector machines.

REFERENCE

1. Baesens, B., S. Höppner and T. Verdonck, 2021. Data engineering for fraud detection. Decis. Support Sys., 10.1016/j.dss.2021.113492
2. Adewumi, A.O. and A.A. Akinyelu, 2017. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. Int. J. Sys. Assur. Eng. Manage., 8: 937-953.
3. Zareapoor, M. and P. Shamsolmoali, 2015. Application of credit card fraud detection: based on bagging ensemble classifier. Procedia Comput. Sci., 48: 679-685.
4. Fiore, U., A.D. Santis, F. Perla, P. Zanetti and F. Palmieri, 2019. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf. Sci., 479: 448-455.
5. Bahnsen, A.C., A. Stojanovic, D. Aouada and Ottersten, 2014. Improving credit card fraud detection with calibrated probabilities. Proceedings of the fourteenth SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014. 2014 Society for Industrial and Applied Mathematics. 677-685.
6. Sahin, Y and E. Duman, 2011. Detecting credit card fraud by decision trees and support vector machines. Int. multiconference Eng. Comput. Sci., 1: 442-447.
7. Bhattacharyya, S., S. Jha, K. Tharakunnel and J.C. Westland, 2011. Data mining for credit card fraud: A comparative study. Decision Support Syst., 50: 602-613.
8. Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Discov., 2: 121-167.