JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Shubham Bhagwat, Vedant Deshpande, Kunal Inglunkar

***Abstract*—**
The rapid growth in E-Commerce industry has led to an exponential increase in the use of credit cards for online purchases and consequently they have been surging in the fraud related to it. Nowadays, the development of technology is rapidly increasing, including the credit card fraud. The credit card fraud (CCF) is one of the problems our banking system is facing today. Fraudsters used many methods to attack the customer. The growth in electronic transactions has resulted in a greater demand for fast and accurate user identification and authentication. Conventional method of identification based on possession of pin and password are not all together reliable. Higher acceptability and convenience of credit card for purchases have not only given personal comfort to customers but also attracted a large number of attackers.
In recent years, for banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used.
The three techniques are applied for the dataset and work is implemented in python language. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy and error rate. The comparative results will show which
Methodologies are efficient.

Keywords: Fraud detection, Credit card, Logistic regression, Decision tree, Random forest.

I. INTRODUCTION

In recent years, the prevailing data mining concerns people with credit card fraud detection model based on data mining. Since our problem is approached as a classification problem, classical data mining algorithms are not directly applicable. This project is to propose a credit card fraud detection system using supervised learning algorithm. supervised algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. Credit card is the most popular mode of payment. As the number of credit card users is rising world-wide, the identity theft is increased, and frauds are also increasing. In the virtual card purchase, only the card information is required such as card number, expiration date, secure code, etc. Such purchases are normally done on the Internet or over telephone.

To commit fraud in these types of purchases, a person simply needs to know the card details. The mode of payment for online purchase is mostly done by credit card. The details of credit card should be kept private. To secure credit card privacy, the details should not be leaked. Different ways to steal credit card details are phishing websites, steal/lost credit cards, counterfeit credit cards, theft of card details, intercepted cards etc. For security purpose, the above things should be avoided. In online fraud, the transaction is made remotely and only the card's details are needed. A manual signature, a PIN or a card imprint are not required at the purchase time. In most of the cases the genuine cardholder is not aware that someone else has seen or stolen his/her card information.

The simple way to detect this type of fraud is to analyse the spending patterns on every card and to figure out any variation to the "usual" spending patterns. Fraud detection by analysing the existing data purchase of cardholder is the best way to reduce the rate of successful credit card frauds. As the data sets are not available and also the results are not disclosed to the public. The fraud cases should be detected from the available data sets known as the logged data and user behaviour. At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence.

Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:

- Credit Card Frauds: Online and Offline
- Card Theft
- Account Bankruptcy
- Device Intrusion
- Application Fraud
- Counterfeit Card
- Telecommunication Fraud

Some of the currently used approaches to detection of such fraud are:

- Artificial Neural Network
- Fuzzy Logic
- Genetic Algorithm
- Logistic Regression
- Decision tree
- Support Vector Machines
- Bayesian Networks
- Hidden Markov Model
- K-Nearest Neighbour

II. EXISTING WORK

Shen, Y. Tal (2007) demonstrate the efficiency of classification models to credit card fraud detection problem and the authors proposed the three classification models i.e., decision tree, neural network and logistic regression. Among the three models' neural network and logistic regression outperforms than the decision tree. [1]

M.J. Islam et al (2007) proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbour classifier is implemented and applied to the dataset for credit card system. [2]

Sahin and E. Duman (2011) has cited the research for credit card fraud detection and used seven classification methods took a major role. In this work they have included decision trees and SVMs to decrease the risk of the banks. They have suggested Artificial Neural networks and Logistic Regression classification models are more helpful to improve the perfor- mance in detecting the frauds. [3]

Y. Sahin, E. Duman (2011) has cited the research, used Artificial Neural Network and Logistic Regression Classification and explained ANN classifiers outperform LR classifiers in solving the problem under investigation. Here the training data sets distribution became more biased and the distribution of the training data sets became more biased and the efficiency of all models decreased in catching the fraudulent transactions. [4]

The IBM Proactive Technology Online (PROTON) open-source tool to cope with uncertainty. The inclusion of uncertainty aspects impacts all levels of the architecture and logic of an event processing engine. The extensions implemented in PROTON include the addition of new built-in attributes and functions, support for new types of operands, and support for event processing patterns to cope with all these. The new capabilities were implemented as building blocks and basic primitives in the complex event processing programmatic language. This enables implementation of event-driven applications possessing uncertainty aspects from different domains in a generic manner [1]
Most of the techniques based on Artificial Intelligence, Fuzzy logic, neural network, logistic regression, naïve Bayesian, Machine learning, Sequence Alignment, decision tree, Bayesian network, meta learning, Genetic Programming etc., these are evolved in detecting various credit card fraudulent transactions.[3]

III.PROPOSED METHODOLOGY

Problem Definition

The Credit Card Fraud Detection Problem includes modeling past credit card transactions with the knowledge of the ones that turned out to be fraud. Here we solve the classification problem by using various encoding and decoding techniques to balance the dataset.

Proposed Algorithm

Algorithm steps for finding the best Algorithm:

Step 1: Import the dataset.
Step 2: Convert the data into dataset format.
Step 3: Do Random Sampling.
Step 4: Decide the amount of data for training data and testing data.
Step 5: Give 70% data for training data and remaining data as testing data (i.e. 30%).
Step 6: Assign train dataset to the models.
Step 7: Apply the algorithm across different algorithms and create the model.
Step 8: Make predictions for test dataset for each algorithm.
Step 9: Calculate accuracy of each algorithm using confusion matrix.

Algorithms used:

1)Logistic Regression
Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).
logistic regression equation:

y = e^ (b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x).
Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.
The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

2)XGBoost
XGBoost falls under the category of Boosting techniques in Ensemble Learning.

In boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors.
Hence, the tree that grows next in the sequence will learnfrom an updated version of the residuals.
In contrast to bagging techniques like Random Forest, in which trees are grown to their maximum extent, boosting makes use of trees with fewer splits

The boosting ensemble technique consists of three simple steps:

An initial model F0 is defined to predict the target variable y. This model will beassociated with a residual (y – F0)
A new model h1 is fit to the residuals from the previous step
Now, F0 and h1 are combined to give F1, the boosted version of F0. The meansquared error from F1 will be lower than that from F0:

$$F1(x) < -F0(x) + h1(x)$$

To improve the performance of F1, we could model after the residuals of F1 and create anew model F2:

$$F2(x) < -F1(x) + h2(x)$$

This can be done for *'m'* iterations, until residuals have been minimized as much as possible:

$$Fm(x) < -Fm - 1(x) + hm(x)$$

Here, the additive learners do not disturb the functions created in the previous steps. Instead, they impart information of their own to bring down the errors.

3) Random Forest

- First, start with the selection of random samples from a given dataset.
- Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- In this step, voting will be performed for every predicted result.
- At last, select the most voted prediction result as the final prediction result.

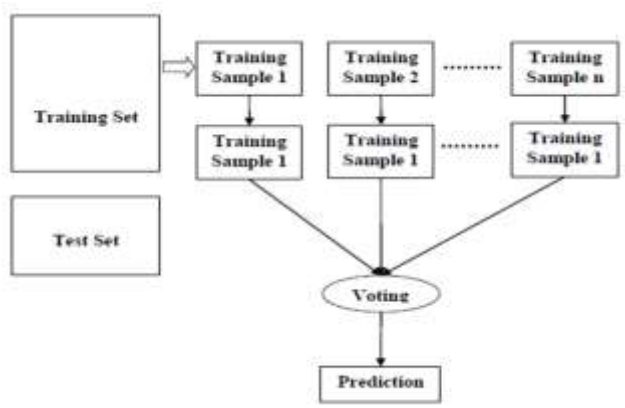

Figure 1. General Algorithm Steps

AdaBoost is a machine learning algorithm. Mainly developed for binary classification.This algorithm is used to boost the performance of decision tree.
For AdaBoost, each instance in the training dataset is weighted. Initial weight is setTo:
Weight (xi) = (1/n) Where, xi – ith training instance n – Number of training instance
This algorithm mainly for classification rather than regression. So that AdaBoostalgorithm is used in fraud detection because this classify the transaction which transactions are fraudulent and non-fraudulent

5)Catboost

CatBoost, short for Category Boosting, is an algorithm that is based on decision trees and gradient boosting like XGBoost, but with even better performance.
CatBoost starts by shuffling the data, creating ***"permutations"***.
For each, it assigns a "default" value for each class to the first few examples

Next, it calculates the value in each new row by looking at previous examples with thesame class, and counting the number of positive labels, then performing a calculation.
This captures additional valuable information, avoids ***"sparsity"***, and speeds upcomputation.
Then the model proceeds by building ***"symmetric binary trees"*** for each permutationof the data.
To avoid overfitting, CatBoost builds new models at each step (n), by shuffling therows and looking at n^2 previous examples.

6)LightGBM

Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred toas boosting.
Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "*gradient boosting*," as the loss gradient is minimized as the model is fit, much like a neural network.

IV.General Framework

The proposed techniques are used in this system, for detecting the frauds in credit card system. The comparison is made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions. The Fig 5.1 shows the architectural diagram for representing the overall system framework.

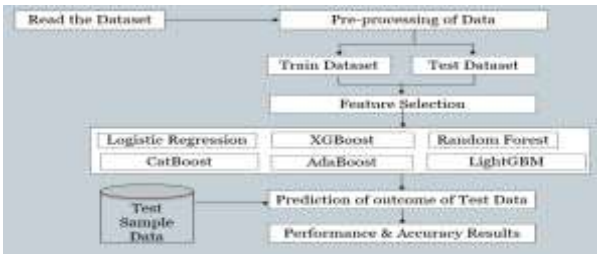

Figure 2. Framework

Steps for Designing the Architecture:

1: Read the dataset.
2: Random Sampling is done on the data set to make it balanced.
3: Divide the dataset into two parts i.e., Train dataset and Test dataset.
4: Feature selection are applied for the proposed models.
5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

6: Then retrieve the best algorithm based on efficiency for the given dataset.

- **Dataset:** In this paper credit card fraud detection dataset was used, which can be downloaded from Kaggle. This dataset contains transactions, occurred in two days, made in September 2013 by European cardholders.
- The dataset contains 31 numerical features. Since some of the input variables contains financial information, the PCA transformation of these input variables were performed in order to keep these data anonymous. Three of the given features weren't transformed.
- Feature "Time" shows the time between first transaction and every other transaction in the dataset. Feature "Amount" is the amount of the transactions made by credit card. Feature "Class" represents the label and takes only 2 values: value 1 in case fraud transaction and 0 otherwise.
- **Divide the dataset**: The dataset is divided into trained data set and test data set. 70% of the data set is under training and the remaining 30% is under testing. Here we are using some supervised machine learning algorithms.
- **Test data:** After training is done on the dataset then testing process take place.
- **Outcome for test data**: We will get the respective results for each algorithm and performance is displayed in graphs.
- **Accuracy results:** Finally results of each algorithm are shown with accuracy and the best algorithm is identified.

V.EXPIREMENT SETTING

A) Assumptions and dependencies:

- Login must be done by the customer.
- User must have the knowledge of English.
- Bank Server should be protected.
- Bank Staff can access all the accounts.

B) System Features:

Administration and Coordinator Module**:**

The administration module and coordinator module will include the following features:

Login – Login page for the administrator and coordinator. All admin and coordinators are identified by the username, password. Administrator and coordinator can create new examination maintain question banks.

Logout – By clicking this link admin user and coordinator logged out from this site all user's session will reset to default value.

C) Software interfaces

This is the software configuration in which the project was shaped. The programming language used, tools used, etc. are described here.

Operating System	Windows 7
Front End	Python
Tool	Jupyter Notebook

VI. DATASET

This dataset is used to detect the credit card fraud detection. This is a classification problem. This is an imbalanced dataset based on target variable. So In this Project, we will use encoding and decoding techniques to balanced dataset.

VII.RESULT AND DISCUSSION

AUC score for all six algorithms were compared and the best one was selected
AUC score for logistic regression is 0.90 for random forest it is 0.85, for xgboost it is 0.97, for adaboost 0.83, for catboost 0.86, for lightGBM it is 0.94
From the AUC scores it is clear xgboost has the best AUC score

CONCLUSION

From this project, Machine learning techniques like Logistic regression, Decision Tree and Random forest were used to detect the fraud in credit card system. By comparing all the three method, we will find the best solution based on performance, efficiency.

REFERENCES

[1]. Fabiana Fournier, Ivo carriea, Inna skarbovsky, The Uncertain Case of Credit Card Fraud Detection, the 9th ACM International Conference on Distributed Event Based Systems (DEBS15) 2015.
[2]. Yashvi Jain, Namrata Tiwari, Shripriya Dubey, Sarika Jain, A Comparative Analysis of Various Credit Card Fraud Detection Techniques, Blue Eyes Intelligence Engineering and Sciences Publications 2019.
[3]. Dinesh L. Talekar, K. P. Adhiya, Credit Card Fraud Detection System-A Survey, International journal of modern engineering research (IJMER) 2014.
[4] A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.
[5] M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. SidAhmed, "Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2007.
[6] Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.