Credit Card Fraud Detection System

Kartik Madkaikar, Manthan Nagvekar, Preity Parab, Riya Raikar, Supriya Patil

Abstract: Credit card fraud is a serious criminal offense. It costs individuals and financial institutions billions of dollars annually. According to the reports of the Federal Trade Commission (FTC), a consumer protection agency, the number of theft reports doubled in the last two years. It makes the detection and prevention of fraudulent activities critically important to financial institutions. Machine learning algorithms provide a proactive mechanism to prevent credit card fraud with acceptable accuracy. In this paper Machine Learning algorithms such as Logistic Regression, Naïve Bayes, Random Forest, K- Nearest Neighbor, Gradient Boosting, Support Vector Machine, and Neural Network algorithms are implemented for detection of fraudulent transactions. A comparative analysis of these algorithms is performed to identify an optimal solution.

Keywords: Error Back Propagation Algorithm (EBPA), K-Nearest Neighbor (KNN), Support Vector Machine (SVM).

I. INTRODUCTION

Objective – The objective of this paper is to detect fraudulent credit card transactions over non-fraudulent transactions and to use machine learning algorithms to predict fraud efficiently and accurately. There are different types of credit card fraud based on the nature of fraudulent activities such as card getting stolen, obtaining cards using false information, individuals using credit cards while being unable to pay debts, bank employees stealing card details to use it remotely, individual using skimming devices to hack credit card details, etc. Vaishnavi Nath, et al. [4] used two methods under random forests namely Random-tree-based random forest and classification and regression tree (CART)-based to train the behavioral features of normal and abnormal transactions. Random forest algorithm performed better on a small dataset, but imbalanced data reduced the accuracy. Aleskerov, et al. [8] implemented CARDWATCH, a database mining system based on a neural network learning module. The system trains a neural network with the past data of a particular customer, data used to process the current spending behavior and detect anomalies. However, it is unrealistic to assume that with this procedure every fraudulent can be detected, because a customer may want to buy an unusual product, or the card number thief may fit into the customer's profile. In this paper, features are selected by performing different feature selection methods like Select-K-Best, Feature Importance, Pearson's Correlation, Mutual Information, Step Forward Selection, Recursive Feature Elimination, Exhaustive Feature Selection. The classification is implemented using various Machine Learning algorithms such as Logistic Regression, Naïve Bayes, Random Forest, K- Nearest Neighbor, Gradient Boosting, Support Vector Machine, and Neural Network algorithms.

Paper is organized as follows: Dataset is presented in section II. Data Pre-processing is presented in section III. Section IV describes Feature Selection. In section V training models are explained. Finally, the result and conclusion are presented in sections VI and VII, respectively.

Manuscript received on July 14, 2021.
Revised Manuscript received on July 18, 2021.
Manuscript published on 30 July, 2021.
* Correspondence Author

Kartik Madkaikar, Student of Bachelor of Engineering, Department of Electronics & Telecommunications Engineering, Padre Conceicao College of Engineering, Verna (Goa), India

Manthan Nagvekar*, Student of Bachelor of Engineering, Department of Electronics & Telecommunications Engineering, Padre Conceicao College of Engineering, Verna (Goa), India

Preity Parab, Student of Bachelor of Engineering, Department of Electronics & Telecommunications Engineering, Padre Conceicao College of Engineering, Verna (Goa), India

Riya Raikar, Student of Bachelor of Engineering, Department of Electronics & Telecommunications Engineering, Padre Conceicao College of Engineering, Verna (Goa), India

Dr. Supriya Patil, Associate Professor, Department of Electronics and Telecommunication Engineering, Padre Conceicao College of Engineering, Verna (Goa), India

II. DATASET

This paper utilizes the dataset provided by revolution analytics for the detection of the fraudulent credit card transaction. Dataset has 51149 legal transactions and 3312 fraudulent transactions. The dataset is divided as 60%, 20% and, 20% in the Train, Valid and Test set, respectively.

III. DATA PREPROCESSING

For efficient implementation of the classification algorithm, data preprocessing is performed before feature selection. Under-sampling is performed to make the dataset balanced to avoid the biasing of the classification algorithm towards the majority class. Feature Selection is implemented on a balanced dataset.

IV. FEATURE SELECTION

Feature selection methods are used to remove unnecessary, irrelevant, and redundant attributes from a dataset that do not contribute to the accuracy of a predictive model or which might reduce the accuracy of the model.

In this paper seven feature selection techniques namely Select-K-best, Feature Importance, Extra tress classifier, Person's correlation, Mutual Information, Step forward selection and Recursive feature elimination are used.

A. Select-K-Best

The scikit-learn library provides the select k best class that is used with a suite of Chi2 statistical tests to select specific numbers of features. Chi2 is used in statics to test the independence of two events and is given by equation (1).

$$X_c^2 = \Sigma \frac{(Si-Ei)^2}{Ei} \qquad (1)$$

C=Degrees of Freedom
S=observed value(s)
E=Expected value(s)
Given the data of two variables, we get observed count S and expected count E. Chi-Square measures how expected count E and observed count S deviates from each other.

B. Feature Importance

Feature importance is a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature at the time of making a prediction. It reduces the number of input features.

In this paper, feature importance is implemented using an extra tree classifier from the decision tree. Extra Trees is similar to Random Forest, it builds multiple trees and splits nodes using random subsets of features, but unlike Random Forest, Extra Tree samples without replacement and nodes are split on random splits.

C. Pearson's correlation

Correlation investigates the relationship between two quantitative continuous variables. A scatter plot of the variables is drawn to check for linearity. If the relationship is not linear the correlation coefficient should not be calculated. A high correlation coefficient value means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related.

$$k = \frac{\Sigma(xi-\bar{x})(yi-\bar{y})}{\sqrt{\Sigma(xi-\bar{x})^2 \, \Sigma(yi-\bar{y})^2}} \qquad (2)$$

k= correlation coefficient,
xi= values of x variable in samples,
yi= values of y variables in samples
x̄= mean of the values of the x variable
ȳ= mean of the values of the y variable

D. Mutual Information

This method is similar to correlation. Mutual information between two variables is a non-negative value which measures dependency between the variable. It is equal to zero if two random variables are independent and higher values mean higher dependency.

E. Step Forward Selection

In Step forward feature selection

- Each individual feature is evaluated and the feature that performs best in selected the algorithm model is selected. The best depends on evaluation criteria.
- Then all possible combinations of that selected feature and a subsequent feature are evaluated, and a second feature is selected, the process is continued until the required predefined number of features are selected.

F. Recursive Feature Elimination

Recursive feature elimination is a wrapper-type feature selection algorithm. It uses different machine learning algorithms at the core of the method. RFE works by searching for a subset of feature by starting with all features in the training dataset and successfully removing features until the desired number remains. In this paper, RFE is implemented with regression.

G. Exhaustive Feature Selection

In exhaustive feature selection, the performance of a machine learning algorithm is evaluated against all possible combinations of the features in the dataset. An Exhaustive search algorithm is the greediest algorithm of all the wrapper methods as it tries all the combinations of the features and selects the best.

Finally, the features that are common to all the implemented feature selection methods are balance, numTrans, creditline, cardholder and dependent variable fraudRisk.

V. TRAINING MODELS

A. Support Vector Machine

Support vector machine is a set of supervised learning methods used for classification, regression, and outlier detection. Different planes (hyperplanes) could be chosen, to separate the data points into two classes. Consider example classification of horse and donkey. The aim is to separate them both based on characteristic data points like the shape of tail, ear, nose which in terms of an algorithm is a plane of separation, the distance between characteristic data points of both classes is termed as margin and the maximum is the better classification. Refer Fig.1 Also, the hyperplane is controlled by the number of features of classification. The hyperplane is just a line if the number of features is two. The hyperplane becomes a multi-dimensional plane if features of classification are two or more, respectively.

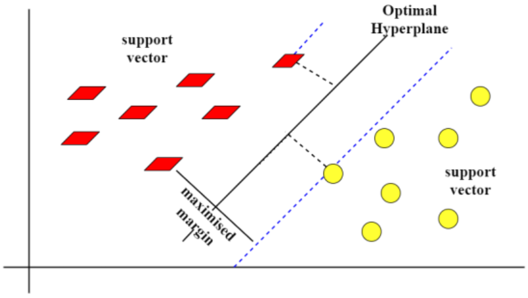

Fig1- SVM

Maximizing the margin distance between them can enable us to classify the input data with more confidence. The Kernel is used for data point separation and to determine the hyperplane. For simplifying the mathematical calculation and for more corrective classification gamma function is used in formula. Following are some of the functions used as per characteristic data input points.

LINEAR KERNEL: The formula of linear kernel is as below usually dot product between any two pairs of observed input values y, yi.

$$K(y, y_i) = \Sigma(y * y_i) \qquad (3)$$

POLYNOMIAL KERNEL: This form of linear kernel type helps us differentiate curved or nonlinear type space with input y, yi respective inputs. Following is the formula for the polynomial kernel:

$$K(y*y_i) = 1 + \Sigma(y*y_i)^{dK(y*y_i)} = 1 + \Sigma(y*y_i)^d \quad (4)$$

Here, D is the degree of the polynomial, which we need to specify manually in the learning algorithm.

RADIAL BASIS FUNCTION (RBF) KERNEL: RBF kernel is mostly used in SVM classification when maps input space is indefinite dimensional space. Following equations (5) and (6) explains it mathematically –

$$K(y*y_i) = e^{\left(-\gamma*\Sigma(y-y_i{}^2)K(y,y_i)\right)} \quad (5)$$

$$= e^{\left(-\gamma*\Sigma(y-y_i^2)\right)} \quad (6)$$

Here, the gamma value ranges 0 - 1.A default value of gamma is 0.1 if not stated manually.

B. K-Nearest Neighbor (KNN)

KNN is a non-parametric classification method that is used to solve classification and regression problems. KNN is termed as a lazy algorithm as it does not do any generalization therefore training process is pretty much fast. Lack of generalization means that the KNN training phase is minimal, or it keeps all the training data. The **M** value should always be chosen as an odd value.

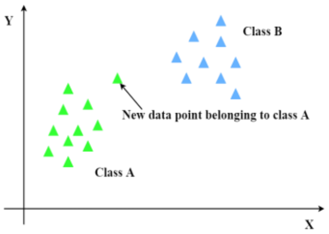

Fig2-Before KNN

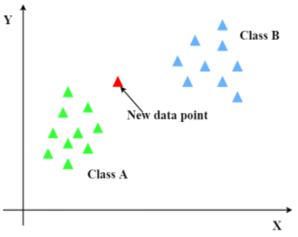

Fig3-After KNN

Algorithm for KNN:

- Input the sample point to be classified as shown in fig2.
- Value for 'M' neighbors is assumed as required.
- Euclidean distances between all the elements of classification is found.
- Depending on the value of 'M' find the M elements with the least Euclidean distances.
- Apply simple majority.
- Plot the sample data point to the majority nearest neighbor class as shown in fig3.

Metric must be defined to find the distances between the sample point and the labeled points of the database. The most common metric used is Euclidean. Some of the others include Euclidean squared, City-block, and Chebyshev:

$$D(v,c) = \{\begin{array}{l}\sqrt{(v-c)^2}Euclidean \\ (v-c)^2 Euclidean^2 \\ |(v-c)| Cityblock \\ Max(|v-c|) Chebyshev\}\end{array} \quad (7)$$

After selecting the value of M, you can make predictions based on the KNN examples using (8). For regression type, KNN predictions are the average of the k-nearest neighbor's outcome.

$$S = \Sigma_{i=1}^{k} S_j \quad (8)$$

Where s_i is the i^{th} case of the examples sample and S is the prediction of the query point and k is the total number of samples.

C. Logistic Regression

Logistic regression is a classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, that is there are only two possible classes. Logistic regression uses an equation as the representation. Input values are combined linearly using coefficient values to predict an output value. A key difference from linear regression is that the output value is modelled as a binary value. Sigmoid activation function is used and is given by:

$$S(X) = \frac{1}{1+e^X} \quad (9)$$

Here, X is the independent variable, and S(X) is the output.
Algorithm for Logistic Regression:

- Input testing data and training data.
- Compute the regression coefficients of training data.
- Using sigmoid function find the relationship between training and testing data.
- Output the object's position.

D. Naïve Bayes

In Naive Bayes for every observation, the probability that it belongs to class 1 or class 2 is determined. Then, the probability of the given conditions is calculated. Naïve Bayes is majorly used in real-world applications which require responding to the user's requests instantaneously. Other common applications include filtering spam from the mails, document classification, or sentiment prediction. The formula of conditional probability:

$$P(E1/E\ 2) = \frac{P(E2/E1)*P(E1)}{P(E2)} \quad (10)$$

A higher probability value is taken as the final class.
Types of Naive Bayes

1) Gaussian Naive Bayes

This type of Naive Bayes is used in the case of continuous variables. It assumes that all the variables have a normal distribution. So, if variables present do not have this property, then it is converted to features having distribution normal.

2) Multinomial Naive Bayes

This is used in case the features represent the frequency. Suppose you have a text document, and you extract all the unique words and create multiple features so that each feature represents the count of the word in the document. In such a case, frequency is a feature. In such a scenario, multinomial Naive Bayes is used. It ignores the non-occurrence of the features. It is known to work well with text classification problems.

3) Bernoulli Naive Bayes

This is used for the binary feature. So, instead of using the frequency of the word if you have discrete features in 1s and 0s that represent the presence or absence of a feature.

In that case, the features will be binary, and we will use Bernoulli Naive Bayes. Also, Bernoulli method will penalize the non-occurrence of a feature, unlike seen in multinomial Naive Bayes.

E. Random Forest

Random forest is an ensemble technique. It is one of the most used algorithms because of its diversity and simplicity. This model uses many decision trees. Each of these decision trees splits out a class of predictions and the class with the majority of the votes becomes our models final output prediction as shown below in fig4. In a random forest, while growing trees instead of searching for the most important features for splitting it looks for the best features among a random subset of features for splitting the nodes. This leads to a wide diversity which will give us a better model. Since there is very little correlation between the different models generated the models generate ensemble predictions that are more accurate than any of the individual predictions. This is because while some trees may be wrong.

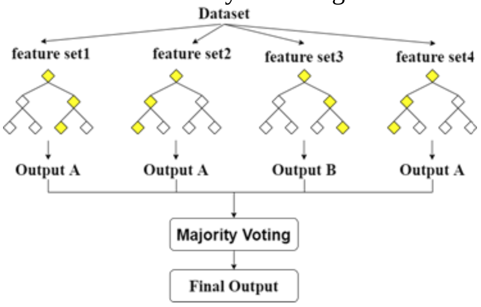

Fig4-Random Forest

Algorithm steps are:

- In this algorithm first, random samples are selected from the given dataset.
- The algorithm creates decision trees for each of the samples selected.
- Algorithm gets the predicted outputs to form each of the decision trees.
- Voting is then performed for all the predicted outputs which were obtained from the decision trees.
- Finally, the result which gets the highest number of votes will be declared as the final predicted result.

F. Error Back Propagation Algorithm

Backpropagation is a supervised learning algorithm for training neural networks. It consists of an input layer a hidden layer and an output layer as shown in fig 5. The principle of the backpropagation approach is to model a given function by modifying the internal weightings of input signals to produce an expected output signal. Error between the system's output and expected output is used to modify its internal state.

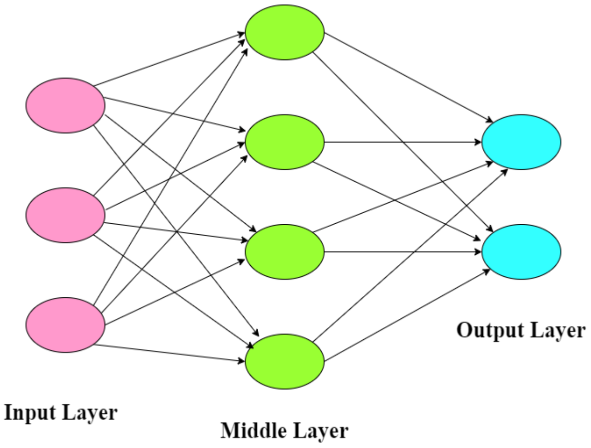

Fig5-EBPA

Each training step of neural network has two main stages

1) *Forward propagation*

Output and hidden layer values are calculated using sigmoid activation function.

$$Output(x) = \frac{1}{1+e^{-x}} \quad (11)$$

2) *Backward propagation (BP)*

The BP stage has the following steps

- Evaluate error signals for each layer using (12).

$$Err_{total} = \Sigma\frac{1}{2}(Target - Output)^2 \quad (12)$$

- Using (13), error gradient is calculated.

$$Err_X = \frac{\partial Err_{total}}{\partial x} \quad (13)$$

- Update layer parameters using the error gradients with an optimization algorithm such a GD

G. Gradient Boosting

Boosting method converts weak learners into strong learners. The gradient boosting algorithm can be most easily explained by first introducing the AdaBoost Algorithm. The AdaBoost algorithm starts by training a decision tree in which each datapoint is given an equal weight. After evaluating the first tree, weights are increased for those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the aim is to improve upon the predictions of the first model. The new model is therefore *Tree 1 + Tree 2*. Classification error is calculated from this new 2-tree ensemble model and grows a third tree to predict the revised residuals. This process is repeated for a specified number of iterations. Predictions of the final ensemble model is the weighted sum of the predictions made by the previous models.

Gradient Boosting trains many models in a slow, additive, and sequential manner. The difference between AdaBoost and Gradient Boosting Algorithm is the way the two algorithms identify the shortcomings of weak learners. While in the case of AdaBoost, the model identifies the shortcomings by using high weight observation point, whereas gradient boosting performs the same by using gradients in the loss function.

The loss function is a measure of indicating how good are model's coefficients fitting the underlying data.

VI. RESULTS

The comparison of accuracy obtained by implementation of preprocessing methods and classification algorithms such as SVM, KNN, Logistic Regression, Naïve Bayes, Random Forest, EBPA, and Gradient Boosting is as shown in Table I.

Table I: Algorithms and their Accuracy

Algorithm	Accuracy
SVM	94.7%
KNN	87%
Logistic Regression	90%
Naïve Bayes	94%
Random Forest	94.9%
EBPA	93.73%
Gradient Boosting	95.9%

VII. CONCLUSION

Fraud detection is a complex process, and it has been proved to be of great help in applications such as identification and prevention of financial frauds like money laundering, credit card thefts, tax evasion, check theft, and embezzlement. Of the various Machine Learning algorithms implemented in this paper, the Gradient Boosting algorithm provides an edge over the other algorithms. Gradient Boosting outperformed the method proposed by Vaishnavi Nath et al. [4] with an accuracy of 95.9%.

REFERENCES

1. An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection SARA MAKKI 1,2, ZAINAB ASSAGHIR2, YEHIA TAHER3, RAFIQUL HAQUE4, MOHAND-SAÏD HACID1, AND HASSAN ZEINEDDINE2.
2. Credit Card Fraud Detection by using ANN and Decision Tree Jasmine A Hudali*, Kamalakshi, K P Mahalaxmi, Namita S Magadum, Prof. Sudhir Belagali.
3. Dataset: http://packages.revolutionanalytics.com/datasets/
4. ICRTAC 2019Credit Card fraud detection using ML algorithms by Vaishnavi Nath Dornadulaa, Geetha Sa.
5. Credit Card Fraud Detection using Various Methods and Techniques by Vasta et al.
6. Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, Fellow, IEEE, and Gianluca Bontempi, Senior Member, IEEE.
7. An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine *ALTYEB ALTAHER TAHA AND SHARAF JAMEEL MALEBAR.*
8. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A neural network based database mining system for credit card fraud detection," in Proc. IEEE/IAFE Computat. Intell. Financial Eng., Mar. 1997, pp. 220–226.
9. Web service-based credit card fraud detection by applying machine learning techniques by Debachudamani Prusti and Santanu Kumar Rath.
10. Fake News Detection with Machine Learning Jayesh Patel, Melroy Barreto, UtpalSahakari, Supriya Patil.
11. Detecting Phishing Websites through Deep Reinforcement Learning by Moitrayee Chatterjee Akbar Siami Namin.
12. Application Of Classification Models On Credit Card Fraud Detection by Aihua Shen, Rencheng Tong, Yaocheng Deng^2.
13. Detecting Credit Card Fraud by ANN and Logistic Regression Yusuf Sahin1 and Ekrem Duman.
14. A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective SamanehSorournejad1, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi.

AUTHORS PROFILE

Kartik Madkaikar, student of Bachelor of Engineering in Electronics & Telecommunications Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India.

Manthan Nagvekar, student of Bachelor of Engineering in Electronics & Telecommunications Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

Preity Parab, student of Bachelor of Engineering in Electronics & Telecommunications Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

Riya Raikar, student of Bachelor of Engineering in Electronics & Telecommunications Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

Dr. Supriya Patil, is an Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. She received B.E (Instrumentation), M.E. (Electronics) degree from Shivaji University and Ph.D. (Electronics) from Goa University. She worked for her Ph.D. in microarray-based cancer classification. She has 25 years of teaching experience with specialization in Signal Processing and Artificial Neural Network. She has authored and co-authored over twenty conference/journal papers in field of Artificial Neural Network