An Approach for Fraud Transaction Detection Using Boosting Model

Neha Jain[1], Amit Ranjan[2]
[1, 2] Dept of Computer Science
[1, 2] SRIST, RGPV-Bhopal, MP, India

Abstract- *Nowadays digitalization gaining popularity because of seamless, easy and convenience use of ecommerce. It became very rampant and easy mode of payment. People choose online payment and e-shopping; because of time convenience, transport convenience, etc. As the result of huge amount of e-commerce use, there is a vast increment in credit card fraud also. Fraudsters try to misuse the card and transparency of online payments. The main aim is to secure credit card transactions; so people can use e-banking safely and easily. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection technique(s) used. This paper investigates the performance of naïve bayes, logistic regression, random forest and Proposed Boosting Model on highly skewed credit card fraud data. This paper is suggesting that a detection model must be available to capture the possible anomalous transactions – a fallback in case the technology will fail. Several classifiers were evaluated during the model creation however only the Boosting Model yielded the highest accuracy. By thorough analysis of these classifiers, it shows that the Boosting Model is more fit in understanding the transaction logs data.*

Keywords- Machine-learning, Credit card, Fraud detection, Fraud transaction, Naive Bayes, Boosting Model.

I. INTRODUCTION

The volume of electronic transaction has risen significantly in recent years, due to the popularization of e-commerce such as online retailers (e.g., Amazon, eBay and Alibaba). Credit/Debit cards are widely used in electronic transaction. Recently, cardless transactions [1] in credit card operations become more and more popular by web payment gateways (e.g. PayPal and AliPay). The global e-commerce market is predicted that it will be worth a staggering US$ 24 trillion by 2019 [2]. However, there has been a simultaneous growth in fraudulent transactions [3] which results in a dramatic impact on the economy. A survey of over 160 companies reveals that the number of online frauds is 12 times higher than that of offline frauds ever year [4]. Since a physical card is not needed in the e-commerce scenario and only the information about the card is enough for a transaction, a fraudster only needs this information for a fraud. For example, after the fraudster steals the account and password of a card from its genuine cardholder, they use them to purchase some goods. Fraudsters usually get card information via a variety of ways: intercepting mails containing newly issued cards, copying and replicating card information through skimmers, or gathering sensitive information through phishing (cloned websites) or from unethical employees of credit card companies [5]. Due to the complexity of the environment and people's background, it is hard to prevent all the genuine cardholder' account from being stolen. The promising way to detect this kind of fraud is to analyze the consuming patterns on every account and to figure out any discrepancies with respect to the "usual" transaction patterns [6].

Currently, there are two kinds of approaches of fraud detection: (1) misuse detection and (2) anomaly detection [7]. The former needs to collect a large database of fraudulent signatures and uses it as a reference to detect the current (mis)use. Such an approach usually has to know the previous cases of fraud in order to obtain the various fraud patterns trends. Various classification methods like neural networks, decision trees, logistic regression and support vector machine have been used in credit card fraud detection [8]. But there are two drawbacks in them: first, it is difficult to obtain all the cases of fraud; furthermore, it fails to detect a new type of fraud that is not recognized via the prior knowledge [9]. In contrast, the latter builds profiles of normal transaction patterns based on historical transaction records, and marks newly observed transactions which deviate from the "average" of "normal" past profiles as potential frauds. Most of the credit card fraud detection methods [10] based on anomaly detection try to extract the historical behavior patterns as rules and compute the similarity between an incoming transaction and these behavior patterns. The main idea of this kind of approach is that people may have personalized transaction habits that depend on their different identities, different incomes, and different motivations and so on. Just pointed out by Adams et al. in [11] there are strong weekly and monthly periodic patterns in a cardholder's transaction behaviors. An aggregated profile which reflect the inherent patterns in time series of transactions are proposed in [11]. A Hidden Markov

Model (HMM) based fraud detection model is proposed by Srivastava et al. [6] and the sequence of operations in credit card transaction processing is modeled by HMM. Sequence alignment method is also used by Amlan et al [10]. In this method the cardholder's historical transaction behavior is presented by a sequence, and sequence alignment is used to determine the similarity of an incoming transaction sequence on a given credit card with the genuine cardholder's past transaction sequences. The objective of this work is to detect the fraudulent financial statements of an organization using an effective and accurate fraudulent financial statement detection model.

II. RELATED WORK

2.1 Classification of load balancing Algorithm

The dirty use of data for e-commerce referred as credit card fraud. Credit card fraud become rampant, as there is increment in credit card transaction. Nowadays, card transaction is not only for the online purchases; it is beyond that in regular purchases also.

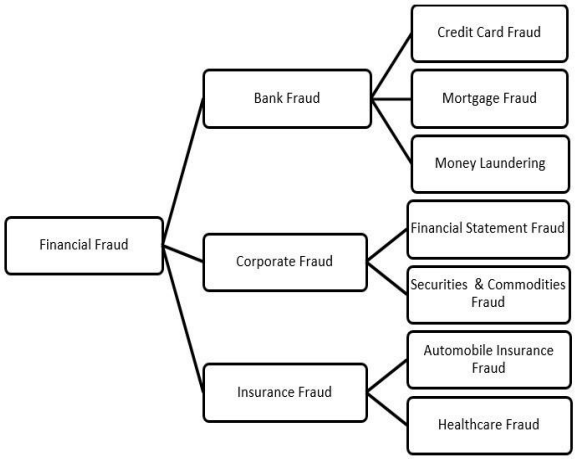

Figure1: Types of Common Fraud [12].

Because of financial fraud by credit card transaction, both merchants and shoppers are suffering from economical loss [13]. It is a very important issue; to solve that issue banks and card manufacture organizations pays significant cost [14]. Credit card fraud detection has been studied mainly on supervised methods, because accurate label information is available in most cases. Machine learning, especially classification models have been tested on transactions to detect fraud through neural networks [18], logistic regressions [19] association rules, modified Fisher discriminant analysis, and decision trees. Since raw input features are not sufficient to detect fraudulent transactions, feature engineering strategies [20] have been proposed. Domain-specific models have also been proposed to address concept-drift and verification latency [17]. An ensemble model, which consists of multiple models, has shown better performance than non-ensemble models such as logistic regression, shallow neural network, support vector machine, and k-nearest neighbor [21]. Several studies have reported that random forest, made by growing many decision trees on randomly generated training samples, achieves the best performance. [16] combined random forests and feed-forward neural networks. In addition, [22] used a model per card account by combining six types of models: decision trees, random forests, Bayesian networks, naive Bayes, support vector machines, and k-nearest neighbours

.Artificial neural networks try to mimic a human's way of processing information. In the 1990s, the shallow neural networks with only one hidden layer containing several nodes were applied to fraud detection problems. In more recent years, deep learning was introduced in several studies [15] [16]. One of the major differences between deep learning and the shallow neural networks is that deep learning models have more complex structure than the shallow models, with more than one hidden layer and more nodes in each layer. [16] Compared various machine learning models including ensemble models and deep feed-forward neural networks. [15] And [16] adopted recurrent neural networks which use a sequence of transactions as an input of the model.

Table below shows some strength and limitations of some data mining techniques used for fraud transaction detections.

Table 1: Strength and limitations of Classification algorithms for Credit Card Fraud Detection.

Model	Strengths	Limitations
Bayesian	Good for binary classification problems. Efficient use of computational resources. Suitable for real-time operations.	Need good understanding of typical and abnormal behaviors for different types of fraud cases.
Trees	Easy to understand and implement. Requires low computational power. Suitable for real-time operations.	Potential of overfitting if the training set does not represents the underlying domain information. Re-training is required for new type of Fraud cases.
Neural Network	Suitable for binary classification problem.	High computational power. Un-suitable for real-time problem. Re-training is required for new types of frauds.
Linear Regression	Provides optimal results when the relationship between independent and dependent variables are almost linear.	Sensitive to outliers. Limited to numeric values only.
Logistic Regression	Easy to implement.	Poor classification performance as compared to other data mining techniques.
Support Vector Machine	Able to solve non-linear classification problems. Requires low computational power. Suitable for real-time operations.	Not easy to process the results due to transformation of the input data.

III. PROPOSED WORK

Figure below is represents overall system architecture. First the credit card dataset is taken from the source than pre-processing is performed on the dataset which includes removal of redundancy, filling empty spaces in columns etc. Dataset is converted into training set by feature engineering and test set. Feature engineering is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. Than classification process is applied to produce results which are evaluated on various parameters.

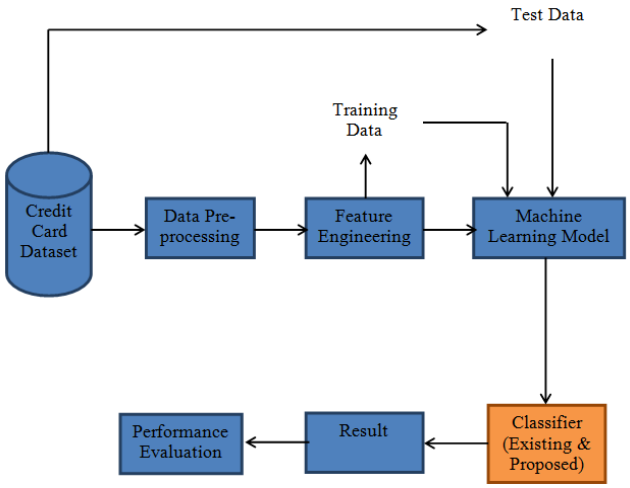

Figure 2 : Proposed architecture

In data pre-processing mainly data cleaning and transformation is performed. Cleaning handles the missing values and outliers if any. It removes noisy, irrelevant, empty, missing values and unwanted data to reduce processing overhead. The data you have selected may not be in a format that is suitable for you to work with. Transformation has been applied to deal with this.

In feature engineering, the features that were obtained in the data were obtained by choosing numerical data such as details of customers, transaction location, business of the merchant and other details that are kept confidential due to the university and credit card issuing authority policy.From a total of 43 attributes that were present in the banks database, a PCA (Principle Component Analysis) was run and 28 important features were obtained.PCA is a dimensionality-reduction technique in which a large number of original variables are condensed into a smaller subset of feature variables.

In the next stage, we need to select a modeling technique to execute this problem. Here, the relationship between dependent and independent variables is derived from the data selected. The question which pattern matches which data is clearly answered by using Machine Learning algorithms. Since the data that is being dealt with is binary and needs to be classified into either fraudulent or non-fraudulent data, proposed machine learning algorithms is used in this case to perform data mining on credit card transactions. After this, result has been evaluated on the basis of accuracy and F-1 score.

For credit card detection various machine learning methods have been used like decision tree, random forest etc. They have some drawbacks. Decision tree has serious disadvantages, including over fitting, error due to bias and error due to variance. Random forests builds each tree independently and it combine results at the end of the process (by averaging or "majority rules")

Proposed system will develop a gradient boosting method for fraud detection. It builds one tree at a time. This additive model (ensemble) works in a forward stage-wise manner, introducing a weak learner to improve the shortcomings of existing weak learners. Gradient boosting combines results along the way.

The Gradient Boosted Tree (GBT) is an ensemble of classification or regression models. It uses forward-learning ensemble models, which obtain predictive results using gradually improved estimations. Boosting helps improve the tree accuracy. The Decision Stump (DS) generates a decision tree with a single split only. It can be used in classifying uneven data sets.

Proposed system will develop a gradient boosting method for credit card fraud detection. Gradient boosting is also a boosting algorithm, which tries to combine the weak learners to form a strong learner. It generates weak learners during the learning process. At each level of the process, the weak learner predicts the values or class label and then calculates the loss, (i.e., the difference between real value and the predicted value). Depending upon the loss, it creates a new weak learner and then the weak learner trains on the remaining errors. This process continues until a certain threshold. This process is called gradient descent optimization problem and therefore this algorithm is called gradient boosting.

In gradient boosting, it trains many models sequentially. Each new model gradually minimizes the loss function (y = ax + b + e, e needs special attention as it is an error term) of the whole system using Gradient

Descent method. The learning procedure consecutively fit new models to provide a more accurate estimate of the response variable.

The principle idea behind this algorithm is to construct new base learners which can be maximally correlated with negative gradient of the loss function, associated with the whole ensemble.

Steps of Gradient Boost algorithm

Step 1: Assume mean is the prediction of all variables.
Step 2: Calculate errors of each observation from the mean (latest prediction).
Step 3: Find the variable that can split the errors perfectly and find the value for the split. This is assumed to be the latest prediction.
Step 4: Calculate errors of each observation from the mean of both the sides of split (latest prediction).
Step 5: Repeat the step 3 and 4 till the objective function maximizes/minimizes.
Step 6: Take a weighted mean of all the classifiers to come up with the final model.

IV. RESULT AND ANALYSIS

The dataset used in this thesis was from the Machine Learning Group of ULB (UniversitLibre de Bruxelles), and it was also released in Kaggle, a community of data scientists and machine learners. It contains the record of credit card transactions made by European cardholders.Dataset contains 30000 transactions. There are 25 attributes. Attributes are limited balance, sex, education, age, amount, bill amount etc.
Confusion matrix is shown below:

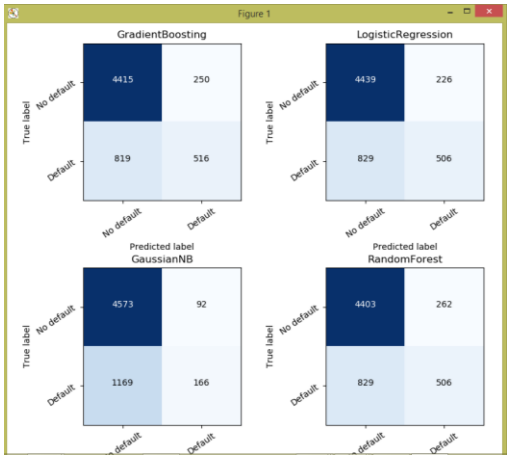

Figure 3:Confusion matrix

Snapshot after execution of the program is shown below:

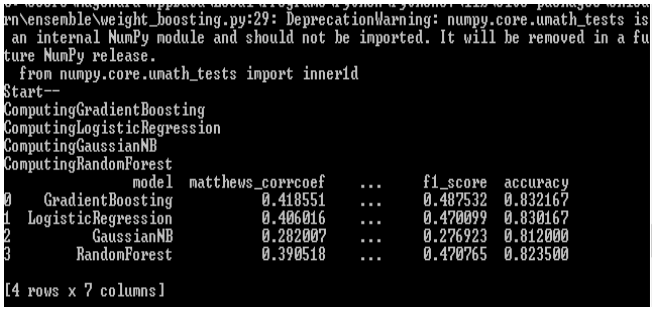

Figure 4:Snapshot of execution.

Table below shows accuracy of different algorithm tested in environment and found that proposed boosting method got good result on each run as compare to other methods..

Table 2: Performance evaluation.

Model	Accuracy (in %)
Gaussian Naive Bayes	81.20
Logistic Regression	83.01
Random Forest	82.35
Proposed	**83.21**

V. CONCLUSION

In this paper we mainly focus on credit card fraud detection in real world. Here the credit card fraud detection is based on fraudulent transactions. Generally credit card fraud activities can happen in both online and offline. But in today's world online fraud transaction activities are increasing day by day. So in order to find the online fraud transactions various methods have been used in existing system. In proposed system we use Gradient Boosting Algorithm (GBA) for finding the fraudulent transactions and the accuracy of those transactions. This algorithm is based on Ensemble type of supervised learning algorithm where it uses multiple random decision trees for classification of the dataset. After classification of dataset a confusion matrix is obtained. The performance of Proposed Algorithm is evaluated based on the confusion matrix.

REFERENCES

[1] S. Gupta S and Johari R. A new framework for credit card transactions involving mutual authentication between cardholder and merchant. In Communication Systems and Network Technologies, pages 22–26. IEEE, 2011.
[2] C. Arun. Fraud: 2016 & its business impact. Technical report, 11 2016.
[3] Vronique Van Vlasselaer, Cristin Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. Apate : A novel approach for automated

credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75:38–48, 2015.

[4] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, and A. K. Majumdar. Credit card fraud detection: A fusion approach using dempstercshafer theory and bayesian learning. Information Fusion, 10(4):354–363, 2009.

[5] Jon T. S Quah and M Sriganesh. Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications An International Journal, 35(4):1721–1732, 2007.

[6] Abhinav Srivastava, Amlan Kundu, Shamik Sural, and Arun Majumdar. Credit card fraud detection using hidden markov model. Dependable & Secure Computing IEEE Transactions on, 5(1):37–48, 2007.

[7] Wen Hua Ju and Yehuda Vardi. A hybrid high-order markov chain model for computer intrusion detection. Journal of Computational & Graphical Statistics, 10(2):277–295, 2004.

[8] V. Dheepa and R. Dhanapal. Behavior based credit card fraud detection using support vector machines. Ictact Journal on Soft Computing, 2(4), 2012.

[9] Z. Zojaji, R. E. Atani, A. H. Monadjemi, et al. A survey of creditcard fraud detection techniques: Data and technique oriented perspective. ArXiv preprint arXiv: 1611.06439, 2016.

[10] Thuraya Razooqi, Pansy Khurana, Kaamran Raahemifar, and Abdolreza Abhari. Credit card fraud detection using fuzzy logic and neural network. In Communications & NETWORKING Symposium, 2016.

[11] Giovanni Montana David J. Weston Niall M. Adams, David J. Hand. Craud detection in consumer credit. Expert Update SGAI, Special Issue on the 2nd UK KDD Workshop, 2(1), 2006.

[12] J. West, M. Bhattacharya, "Intelligent Financial Fraud Detection: A Comprehensive Review", ELSEVIER, Computer & Security, 2016, 47-66.

[13] N. Mahmoudi, E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis", Elsevier Expert System with Application, 2015, pp. 2510-2516.

[14] N. Halvaiee, M. Akbari, "A novel model for credit card fraud detection using Artificial Immune System", Elsevier Applied Soft Computing, 2014, pp. 40-49.

[15] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.

[16] Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Belling, P. (2018). Deep learning detecting fraud in credit card transactions. In Systems and Information Engineering Design Symposium (SIEDS) (pp. 129–134), IEEE.

[17] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S.Waterschoot, G. Bontempi, Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications 41 (10) (2014) 4915–4928

[18] Guo, T., & Li, G.-Y. (2008), Neural data mining for credit card fraud detection. In 2008 International Conference on Machine Learning and Cybernetics (pp. 3630–3634), IEEE volume 7.

[19] Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. Expert Systems with Applications, 39, 12650–12657.

[20] Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using bayes minimum risk. In Proceedings-2013 12th International Conference on Machine Learning and Applications (ICMLA) (pp. 333–338). IEEE Computer Society volume 1.

[21] Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 289–294). ACM.

[22] Kultur Y & caglayan, M. U. (2017). Hybrid approaches for detecting credit card fraud. Expert Systems, 34, e12191.