International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

# Financial Instruments Fraud Detection Using Machine Learning

**Dr. Priti. Subramanium, Bhusawal Evangel Denis Rodrigues**
Dept. of Computer Science,
Shri Sant Gadge Baba College of Engineering and Technology,
Bhusawal

**Abstract- The rapid growth in industries mainly in the e-commerce sector has led to massive use of credit cards for online purchases and consequent surge in the fraud related to it which makes it necessary for banks to keep updating their fraud detection methods. Machine learning approaches are mostly being used to analyze and detect the increasing serious problem of credit card fraud. For discovering these fraud transactions banks make use of various machine learning methodologies. The performance of fraud detection in credit card transactions greatly depends upon the sampling approaches being carried out on the data-sets, selection of variables and detection techniques used in finding out the unauthorized activities. In this project the methods that have been used are logistic regression, neural networks, k-nearest neighbor are used for credit card fraud detection. I have also done oversampling to balance the data set. These techniques are applied on the dataset and output is generated .The performance of the techniques is evaluated for different variables based on sensitivity, accuracy and error rate.**

**Keywords- Fraud detection, Credit card, Machine Learning, fraud transactions**.

## I. INTRODUCTION

A credit card is referred to a card that is assigned to a customer .Usually allowing them to purchase goods and services within the credit limit or withdraw cash in advance. Credit card provides the cardholder an advantage, i.e., it allows their customers time to repay later in a prescribed time, by carrying it to the next billing cycle. For this they need to provide their personal information which makes them easy targets by fraudsters.

Without any kind of risk, a significant amount can be withdrawn without the owner's knowledge .Fraudsters always try to make every fraudulent or unauthorized activity to seem legitimate that it becomes very hard to detect that even a fraud has taken place.

As we know that a certain set of activities that are carried out to prevent money or property from being obtained through false claims is known as fraud detection .The use of credit and debit cards has increased significantly in the last years, unfortunately so has fraud increased on a more higher rate. Because of which, financial frauds has been increasing at a rapid rate. This opens new challenges as new fraud patterns go on emerging, and current fraud detection systems are less successful in preventing these frauds.

With the fear of being caught fraudsters constantly change their strategies or say methods to avoid being detected, something that makes traditional fraud detection tools useless or inadequate. Moreover, machine learning methods as well can be inadequate if they miss to adapt to new fraud strategies, i.e., static models that are never updated would be of no use [2].

## II. LITERATURE SURVEY

The recognized data can be used by banks in order to keep track of their revenue they handle. Usually such kind of applications needs to know previous data and on the basis of this previous data they detect the fraud. There was this other statistical approach is feed forward network in which a certain kind of relationship is found between user data and other parameters to get the result.

Another new emerging technology of Credit card fraud detection is based on the genetic algorithm and scatter search. **Ekrem Du Man, M Hamdi Ozclik** published an approach that was based on genetic algorithm and scattering search.

The Peer group analysis which was made by **David Weston and Whit row** is a good solution regarding credit card fraud detection was a good approach that was based on unsupervised learning and it monitors the behavior over time as well. This technique could be used to find anomalous transaction and help to detect the fraud in time. All these technologies have their pros and cons as well.

As **Linda Delmaire** works on association rule in which a simple method that initially needs large data set in which it can find frequent item set.

As work done by **Lokeh et.al** is on Neural Network that can be applied to both in Supervised as well as

Unsupervised Approach. Since unsupervised approach is little bit more complex but give more optimized results are obtained.

**John T.S Quah & M.Sriganesh** worked on Real time Credit card Fraud detection. They used computational intelligence that works on Self Organising Map. Ekrem et.al combined the genetic algorithm and Scatter search approach that is really helpful to find unauthorized transactions.

**David Weston** provided a good solution to find credit card fraud detection by making use of Peer Group analysis method. So, the main motive of our paper is to represent simple technologies that can detect the fraud as soon as possible and to avoid the loss as well.

## III. PROBLEM STATEMENT

The Credit Card Fraud Detection problem consists modeling past credit card transactions with the knowledge of the ones that turned out to be fraud.

The main and prime objective of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction.

This model is then used to identify whether a new transaction that is being made is genuine or not. Our main objective here is to detect the fraud transaction while minimizing the incorrect fraud.

## IV. PROPOSED SYSTEM

When constructing a credit card fraud detection system, we mainly require a certain set of data related to a certain persons transactions so that we can easily cross check the transaction patterns and techniques.

**Capturing customer spending patterns:** As we know that a single transaction information is not sufficient to detect a fraudulent transaction, since using only the raw features leaves behind important information i.e such as the consumer spending behavior, which is usually used by commercial fraud detection systems. To deal with this, a new set of features were proposed such that the information of the last transaction made with the same credit card is also used to make a prediction.

The main focus is to be able to detect very dissimilar contiguous transactions within the purchases of a customer. The additional set of features include: time since the last transaction, previous amount of the transaction, previous country of the transaction. These features do not take into account consumer's spending habit or behavior other than the last transaction made by a client, this leads to having an incomplete profile of customers. A more compressive way to take into account a customer spending behavior is to derive some features using a transaction aggregation strategy.

This methodology was initially proposed in the derivation of the aggregation features which consists in grouping the transactions made during the last given number of hours: first by card or account number, then by transaction type, merchant group, by country name or other, followed by calculating the number of transactions or the total amount spent on those transactions. When a customer transactions are looked at, there is an important question on how much to accumulate, in the sense that the marginal value of new information may diminish as time passes.

As time passes, information loses its value, in the sense that a customer spending patterns are not expected to remain constant over the years. [7]

**Time features:** When using the aggregated features, there is still some information that is not completely analyzed by those features. In particular we are more interested in analyzing the time of the transaction. The logic when following this way is that a customer is expected to make transactions at similar hours.The issue when dealing with the time of the transaction, i.e specifically, when analyzing a feature such as the mean time of transaction is that it is easy to make the mistake of using the arithmetic mean technique.

## V. SOFTWARE REQUIREMENTS

**1. CPU:**
2*64 –bit 2.8 GHz 8.00 GT/s CPUs

**2. RAM:**
32GB (or 16GB of 1600MHz DDR3 RAM)

**3. Storage:**
300 GB

**4. System Requirement:**
ANACONDA, EXCEL.

## VI. ALGORITHMS

**1. Logistic Regression:**
Logistic Regression is the most popular and most used machine learning algorithms. Logistic regression is not a regression algorithm but a classification algorithm.

The model is trained by using the Logistic Regression algorithm which can be used to describe the relationship among the variables of data whether it is binary, continuous, or categorical. Predictors can also be used to predict if certain things will occur or not in the process.

With the help of this model, we can estimate the probability, if the variable belongs to the class or not. Logistic regression is basically a linear method, but the predictions are transformed using the logistic function.

The impact of this is that we can no longer understand the predictions: if it's a linear combination of the inputs as we can with linear regression technique, for example, continuing on from above, the model can be stated as:

$$p(X) = e^{(b0 + b1*X)} / (1 + e^{(b0 + b1*X)})$$

$$In (p(X) / 1 – p(X)) = b0 + b1 * X$$

**2. Bayes' Theorem:**
The Bayes' Theorem provides a way that we can calculate the probability of a piece of data belonging to a given class, given our prior knowledge. Bayes' Theorem is stated as:

$$P ( class |data) = (P( data| class) * P(class))/P(data)$$

Where P (class |data): probability of class given the provided data.

**3. Neural Networks:**
The concept of neural network is based on the human brain. The Neural network is a concept of deep learning in which it uses different layers to perform computation. The reason it is used because of its ability to provide more accurate results and deep learning. Neural Network uses cognitive learning which is used to create models that can be used to perform certain tasks like data mining, prediction, detection, etc.

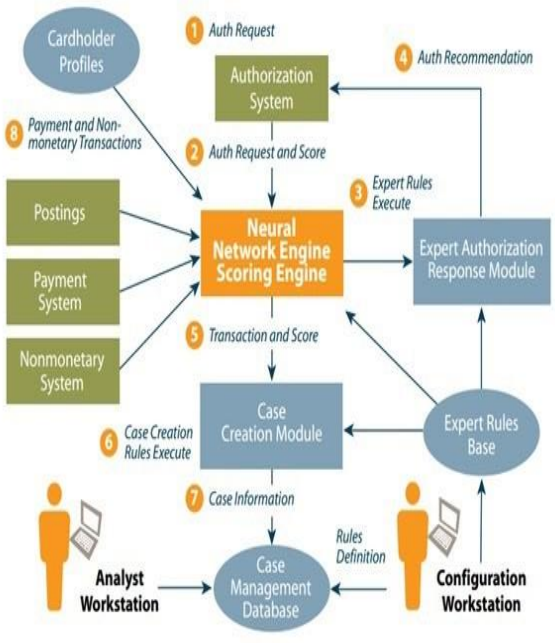


Fig 1. Neural Network.

The neural network is a part of cognitive computing technology where the machine mimics or say copies how the human brain works ,it basically observes patterns. The neural networks are completely adaptive; which makes it able to learn from patterns of legitimate behavior.

These can adapt to the change in the behavior of normal transactions and immediately identify patterns of frauds. The process of the neural networks is extremely immediate and can make decisions in real time.

**4. K-Nearest Neighbour:**
This algorithm is used for both purposes regression as well as a classification problem. KNN classifies the data based on K Nearest neighbors algrothim. It totally depends on labels. In the KNN algorithm classification problem it classifies the data based on its neighbors position.

If the algorithm finds most of the values are of the fraud class, then it declares and separates the dataset in the fraud transaction class. The performance of KNN algorithm is mainly influenced by three main factors: The distance metric used to locate the nearest neighbours, the distance rule used to derive a classification from k-nearest neighbour and the number of neighbours used to classify the new sample.

Amongst the various credit card fraud detection methods of supervised statistical pattern recognition, the K Nearest Neighbour rule is observed to have consistently high performance ability, without a priori assumptions about the distributions, from which the training examples are drawn out.

The KNN based credit card fraud detection techniques require a distance .Or say similar the measure defined between two data instances. In KNN, we mainly classify any incoming transaction by calculating of nearest point to new incoming transaction, then if the nearest neighbour is observed to be fraudulent, then the transaction indicates as a fraud. The value of K is used here as a small, and odd to break the ties (typically 1, 3 or 5), Larger K values can help to reduce the effect of noisy data set. [7]

**5. Decision Tree:**
Combinations of decision trees to improves the results and provide efficiency. Each decision tree checks for different conditions. They are trained on random datasets. And based on the training of the decision trees, each tree gives the probability of the transaction being 'fraud' and 'non-fraud. Then, the model predicts the result accordingly based on the data.[5]

These algorithms are used and regression ,classification all required processes are done on the dataset and output is generated in the form of graphs .With which we can then analyse and find out the transactions that are performed are fraudulent or not.

International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

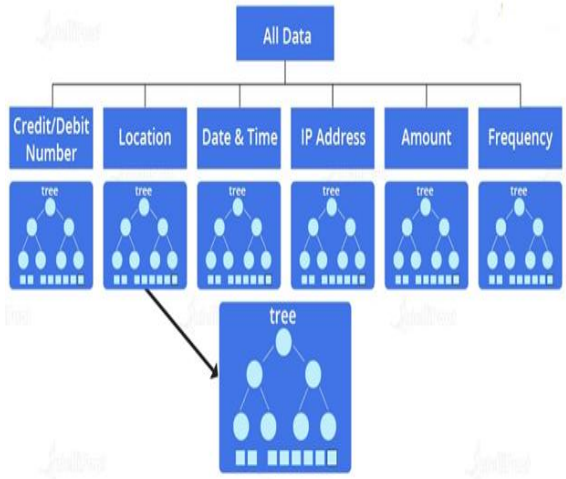


Fig 2. Decision Tree.

## VII. UML DIAGRAMS

**1. Data Flow Diagram:**

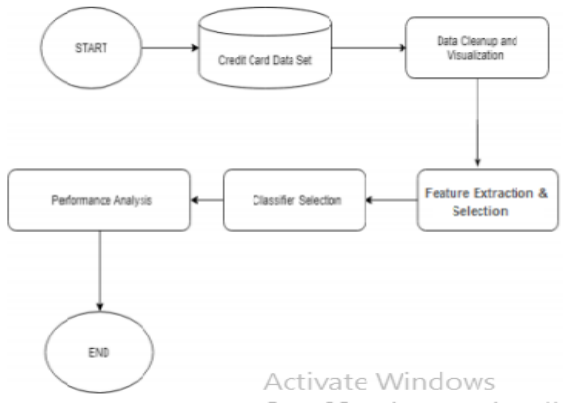


Fig 3. Data Flow.

**2. Use Case Diagram:**

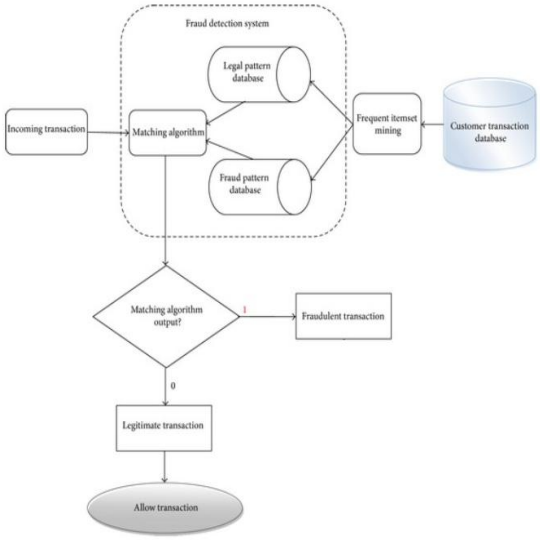


Fig 4. Use Case Diagram.

## VIII. RESULTS

All the experiments are done in Python programming language. The software operating environment is Jupyter notebook which is a part of Anaconda platform. Some of the libraries which have been used include numpy, matplotlib, pandas, sklearn, imblearn. The data sets are splits into training and test data in a ratio of 70:30. This experiment is accomplished in three stages. Firstly, the standard model with SMOTE technique to deal with the imbalanced data set has been used.

From Table 1, we can observe the results of various individual machine learning-based classification models. The outcome of various individual models is depicted in Table 1 indicates that the accuracy of all the models is around 99%. However, the F1 with 0.89 and the MCC score with 0.8900 of RF are slightly high among all the models. Figure 2 shows the score of different evaluation metrics used for the evaluation of several individual classifiers.

It can be seen that the score of all the evaluation metrics for random forest is high in comparison with other classifiers. Inspite of the standard model, AdaBoost along with SMOTE technique have been implemented with these standard models. Therefore, the AdaBoost method is applied. From Table , it can be observed that here also the RF model gives the best result.

Table 1. Results of various individual models.

| ML based classifiers | Accuracy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| Random forest | 0.9996 | 0.95 | 0.83 | 0.89 | 0.8900 |
| Logistic regression | 0.9993 | 0.91 | 0.68 | 0.78 | 0.7876 |
| Naive Bayes | 0.9992 | 0.86 | 0.66 | 0.74 | 0.7479 |
| XGBoost | 0.9995 | 0.92 | 0.82 | 0.87 | 0.8726 |

## ACKNOWLEDGMENT

It gives me great pleasure in presenting the project report on 'Financial Instruments Fraud Detection Using Machine Learning'.I would like to take this opportunity to thank our internal guides Head of Computer Engineering Department, Shri Sant Gadge Baba College of Engineering And Technology, Bhusawal.

Dr. Priti Subramanium Asst. Professor for giving us all the help and guidance needed. I am really grateful for their kind support.

## REFERENCES

[1] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, 14277–14284.

[2] Varmedja, D., et al. (2019). Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE.

[3] Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). IEEE.

[4] Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S., & Aswini, E. (2019). Credit card fraud detection using random forest algorithm. In 2019 3rd International Conference on Computing and Communications Technologies (ICCCT) (pp. 149–153). IEEE.

[5] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) (pp. 1–6). IEEE.

[6] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1–9). IEEE.

[7] Dighe, D., Patil, S., & Kokate, S. (2018). Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1–6). IEEE.

[8] Puh, M., & Brki´c, L. (2019). Detecting credit card fraud using selected machine learning algorithms. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1250–1255). IEEE. 40 Saloni and M. Rout

[9] Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 122–125). IEEE.

[10] Mishra, A., & Ghorpade, C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. In 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1–5). IEEE.

[11] Sisodia, D. S., Reddy, N. K., & Bhandari, S. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 2747–2752). IEEE.

[12] Xie, Y., Liu, G., Cao, R., Li, Z., Yan, C., & Jiang, C. (2019) A feature extraction method for credit card fraud detection.In 2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS) (pp. 70–75). IEEE