IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

Credit Card Fraud Detection on Imbalanced Dataset using Supervised Machine Learning Algorithms

Abhinaya Karthikeyan [1], Manasa Jakkani [2]

Department of Information Technology, B.K.Birla College of Arts Science & Commerce, Kalyan, India[1, 2]

ABSTRACT: Many companies and institutions has been moved their services online in e-commerce for their customers easily accessibility. This has made fraudsters to indulge in committing credit card fraud. Credit card fraud causes many financial losses to institutions and for customers. The credit card fraud detection has 2 major problems first, fraudulent transaction behavior changes constantly and second, there is highly imbalanced data where the number of genuine transactions are higher than the fraudulent transactions. In this paper we implement different machine learning algorithms on an imbalanced dataset. To detect fraud transactions and to find best accuracy we used four machine learning algorithms K-nearest neighbor, Random Forest, Logistic Regression and Decision Tree with various techniques. The performance of technique is evaluated based on accuracy and precision. The work is implemented in python. In this research paper, we also compare that which algorithms provide much efficient accuracy for fraudulent transactions of Imbalanced dataset.

KEYWORDS: Imbalanced dataset, Supervised Machine learning algorithms, Random Forest, Logistic Regression, K-Nearest neighbor, Decision tree.

I. INTRODUCTION

Since the beginning of the digital era, the number of cashless transactions is at its peak point and it is most likely to increase in the future[1]. Due to increasing popularity of cashless transactions, one of the most common frauds is credit card frauds. According to Nilson Report, the global credit card fraud losses reached $16.31 billion in 2014 and it is estimated that it will exceed $35 billion in 2020[1]. While that is an advantage and provides ease of use for customers, but it also creates opportunities for fraudsters. As the loss is quite major, hence there are a number of researches to decrease the causalities created by credit card fraud. Credit card fraud is a kind of unauthorized activity or theft to make payment using credit card in an electronic payment system as a fake source of fund. Credit card fraud refers to the situation where fraudster uses credit card for their needs while owner of that credit card is not aware of that[2]. Although there is a tremendous volume increase in credit card transactions, the amount of frauds is proportionally the same or has decreased due to sophisticated fraud detection systems. However, fraudsters are constantly coming up with new ways to steal information. There are two types of credit card frauds. One is theft of physical card, and other one is stealing sensitive information from the card, such as card number, cvv code, type of card and other. By stealing credit card information, a fraudster can broach a large amount of money or make a large amount of purchase before cardholder finds out. Because of that, companies use various machine learning methods to recognize which transactions are fraudulent and which are not[2]. The purpose of this paper is to analyse various machine learning algorithms and see accuracy of these algorithms with respect to credit card transaction dataset, such as Logistic Regression (LR), Random Forest (RF), KNN and Decision Tree in order to determine which algorithm is most suitable for credit card fraud detection.

II. CREDIT CARD FRAUD DETECTION

Credit card fraud affect the organization by financial losses and due to this individual user also affected if the information of credit card get steal[4]. So it is important to detect the fraudulent transactions which classify a transaction into fraud or non-fraud. Many techniques have been developed for credit card fraud detection like Artificial Intelligence & Machine Learning and also based on locations. In this paper, we focus on Machine Learning technique; basically it finds the accuracy of the various algorithms on imbalanced dataset. Machine learning is this generation's

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

solution which replaces such methodologies and can work on large datasets which is not easily possible for human beings. Machine learning techniques fall into two main categories; supervised learning and unsupervised learning[5].

Supervised Algorithms

Supervised learning is a machine learning method in which models are trained using labelled data. In supervised learning, models need to find the mapping function to map the input variable with the output variable. This model takes direct feedback to check if it is predicting correct output or not. This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). The goal of supervised learning is to train the model so that it can predict the output when it is given new data. It needs supervision to train the model. Using these set of variables, we generate a function that map inputs to desired outputs. Supervised learning can be categorized in Classification and Regression problems. The training process continues until the model achieves a desired level of accuracy on the training data[6]. Examples of Supervised Learning: Logistic Regression, Decision Tree, KNN, Logistic Regression, Naive Bayes, etc.

Unsupervised Algorithms

Unsupervised learning algorithms are trained using unlabeled data. This model finds the hidden patterns in data. In unsupervised learning, only input data is provided to the model. The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset. Unsupervised learning does not need any supervision to train the model. Unsupervised Learning can be classified in **Clustering** and **Associations** problems. These algorithms can be used for those cases where we have only input data and no corresponding output data[6]. Unsupervised learning algorithms may give less accurate result as compared to supervised learning algorithms.
Some of the unsupervised learning algorithms are K-means clustering, Hierarchal clustering, Anomaly detection, Neural Networks, Principle Component Analysis. , Independent Component Analysis, Apriori algorithm.

We are going to compare the following four algorithms and check which supervised algorithm is more accurate on imbalanced dataset:

1. Decision Tree

Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. With the strategy of separating and resolving, decision tree usually separates the complex problem into many simple ones and resolves the sub-problems through repeatedly using, which is a data mining method to discover training several kinds of classifying knowledge constructing decision tree. The core of decision tree model is how to construct a decision tree with high precision and small scale[7]. The decision tree is a table of tree shape with connecting lines. Each node is either ramification node followed with more nodes or only one leaf node signed by classification. Decision tree method has many advantages. The first is the high flexibility that it is a non-parameter method without any assumption for the data distribution, and the second is the good haleness. Besides, it is explainable, which is also the reason of its wide utilization.

2. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing[8].

3. KNN

Here third algorithm which we are going use is k- Nearest Neighbor model. KNN(K-Nearest neighbour) is one the simplest but most effective model. This method was introduced for the first time by Aha, Albert and Kibler in 1991.In this algorithm, the class label of the test datasets on the basis of the class labels of the neighboring training data elements. The similarity between two elements is measured using Euclidean Distance. It is also known as an Instance learning or Lazy model. The value of 'k' is calculated which actually the number of its nearest neighbors that have to

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

be considered. A suitable value for 'k' should be chosen. An appropriate distance metric is also a requirement. It is a generalization of the Euclidean and Manhattan distance. Mathematically, it is can be represented as:

$$d(x^{(i)},x^{(j)}) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p}$$

4. Logistic Regression

It is basically a statistical model which makes use of a logistic function to model a binary dependent variable. This model is mainly used where there is a chance of occurrence of a binary classification issue. It works well on linearly separable classes. The odds ratio is one concept using which we can also define the logit function. It is the probability of an event occurring.

Odds Ratio = p/(1 — p)

Where, p = probability of the positive event. The logit function is the logarithm of the odds ratio. It takes input in the range of [0,1] and transforms them to values over the real-number range[9]. The logit function can be defined as follows:

Logit (P) = $\log \frac{p}{1-p}$

In this model, the sigmoid function is also used effectively

$$\emptyset(z) = \frac{1}{1 - e^{-z}}$$

This model is basically an ensemble classifier, i.e. a combining classifier that uses and combines many decision tree classifiers. The main agenda behind using multiple trees is to be able to train the trees enough, such that, contribution from each of them comes in the form of a model. After the generation of the tree, the output is combined through majority. It uses multiple decision trees so that, the dependence of each of them is on a particular dataset, possessing similar distribution throughout the tree. This particular model has the quality of efficiently balancing errors in a class population of unbalanced data sets. It can be used to solve both classifications as well as regression problems. Logistic regression (LR) is useful for situations in which we want to be able to predict the presence or absence of a characteristics or outcome based on values of a set of predictor variables[10].

III. METHODOLOGY

Dataset

In this project we use imbalanced dataset. An imbalanced dataset is a dataset where classes are distributed unequally[11]. The dataset if credit card fraud detection is taken from kaggle. The dataset contains transactions occurred in 2 days by European cardholders in September 2013. There are total 284,807 transactions in which 492 i.e., 0.172% transactions are fraudulent. The percentage of fraudulent transaction is extremely low, so the dataset is highly imbalanced.

The dataset contains 31 numerical features. Since some of the input variables contain financial information, the PCA transformation of these input variables was 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 the 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 of fraud transaction and 0 otherwise. Dataset contains 284,807 transactions where 492 transactions were frauds and the rest were genuine[2]. Considering the numbers, we can see that this dataset is highly imbalanced, where only 0.173% of transactions are labeled as frauds. Since distribution ratio of classes plays an important role in model accuracy and precision, preprocessing of the data is crucial.

Pre processing

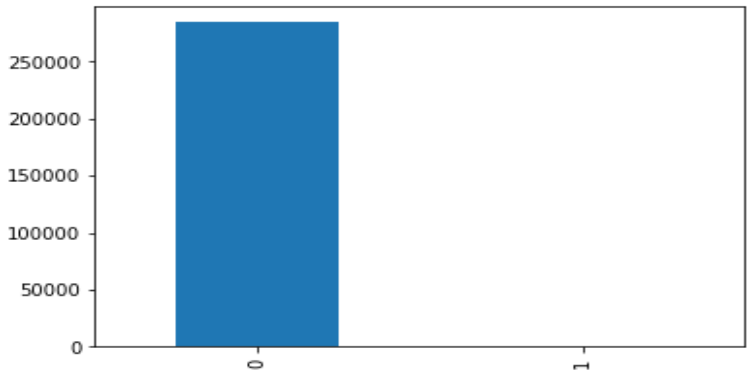
Feature selection is a fundamental technique, which selects the variables that are most relevant in the given dataset. Carefully choosing appropriate features and removing the less important one can reduce over fitting, improve accuracy and reduce training time. Visualization techniques can be helpful in that process. Feature selector tool is by using this tool it has been determined which features are the most important. Furthermore, features that do not contribute to the cumulative importance of 95% were removed. After the feature selection technique, 27 features were selected for additional experiment. Machine learning algorithms have trouble learning when classification categories are not approximately equally distributed. Considering given data is highly imbalanced, it is necessary to perform some kind of

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

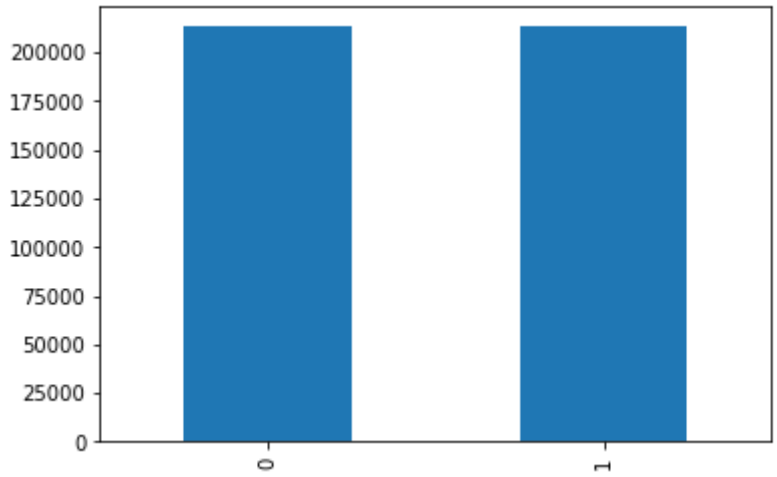
|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

balancing, so that model can be efficiently trained[12]. Frequently used methods for adjusting the class distribution include undersampling the majority class, oversampling the minority class, or combination of those two[2]. Synthetic Minority Oversampling Technique (SMOTE) is a popular oversampling method that has proven useful when used on imbalanced dataset SMOTE was proposed method to improve random oversampling (Fig. 1). Many machine-learning algorithms expect the scale of the input. Taking into account that values of time and amount are highly varying, scaling is done in order to bring all features to the same level of magnitudes. Figure 1. Class distribution before and after sampling the experiment system environment is Windows 10 operating system, and the software operating environment is Jupyter, scientific python development environment, which is part of the Anaconda platform. Used libraries include: numpy, pandas, matplotlib, sklearn and imblearn. Previously mentioned algorithms used in the experiment are described in the following section.



Before Transaction Class Distribution



After Transaction Class Distribution

Experiments

In our analysis, we used supervised machine learning algorithms, on a single and imbalanced dataset, with transactions previously specified fraudulent or not. As we mentioned above four algorithms: Decision tree, Random Forest, K-Nearest Neighbor and Logistic Regression. We have applied these algorithms on the trained dataset and found the desired result

Various parameters for measuring performance have been computed. In this data analysis python programming language is being used to implement different classifier algorithms of machine learning to confirm credit card fraud

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

from those in the dataset. During implementation, the data set has been divided into two categories of 70 % for training and 30 % for testing[13].

Decision Tree algorithm uses a recursive and descending approach to create branches without going back. The idea behind this technique is to create a tree that recursively splits the learning dataset into subsets based on a single attribute. Knn based credit card fraud detection requires two main things to estimate namely, the distance or the measure of the similarity between instances of two data i.e. fraudulent and genuine. The logistic regression model describes relationship between predictors that can be continuous, binary, and category. Based on some predictors we predict whether something will happen or not. We estimate the probability of belonging to each category for a given set of predictors. .Random forests function as a large collection of decision trees. It is considered as an ensemble-learning model for prediction and classification .This algorithm was able to give a high performance than other algorithms.

Algorithms	Accuracy	Precision
Logistic Regression	0.98823	0.98833
Decision Tree	0.99775	0.99810
Random Forest	0.99948	0.99969
KNN	0.96446	0.98969

Table.1

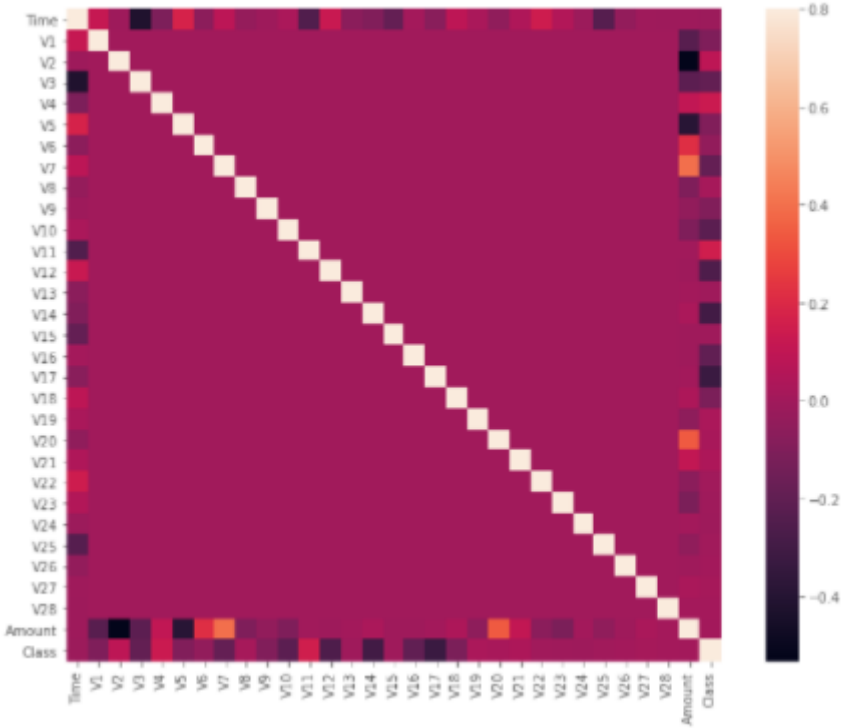


Fig.3

IV. CONCLUSION

Credit card fraud has without hesitation an expression of criminal deception. It represents a very serious business problem. These frauds can lead to huge losses, both business and personal. Because of that, companies invest more and more money in developing new ideas and ways that will help to detect and prevent frauds.

The main aim of this paper was to compare certain supervised machine learning algorithms for detection of fraudulent transactions and to find the accuracy of the algorithms. Hence, comparison was made and it was established that Random Forest algorithm gives the best results i.e. gives efficient accuracy compared to other algorithms which we have used in our paper i.e. KNN, Decision Tree and Logistic Regression for our dataset. This was established using different metrics, such as accuracy and precision. Further research should focus on different machine learning

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008156 |

algorithms such as genetic algorithms, and different types of stacked classifiers, alongside with extensive feature selection to get better results.

REFERENCES

[1]. Y. Yazici, "Approaches to Fraud Detection on Credit Card Transactions using Artificial Intelligence Methods," in *Computer Science & Information Technology*, Jul. 2020, pp. 235–244. doi: 10.5121/csit.2020.101018.
[2]. O. S. Yee and S. Sagadevan, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," vol. 10, no. 1, p. 6.
[3]. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, Mar. 2019, pp. 1–5. doi: 10.1109/INFOTEH.2019.8717766.
[4]. A. U. S. Khan, N. Akhtar, and M. N. Qureshi, "Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm," p. 9.
[5]. A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, Jan. 2019, pp. 488–493. doi: 10.1109/CONFLUENCE.2019.8776942.
[6]. "Supervised vs Unsupervised Learning - Javatpoint." https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning (accessed Aug. 28, 2021).
[7]. A. Shen, R. Tong, and Y. Deng, "Application of Classification Models on Credit Card Fraud Detection," in *2007 International Conference on Service Systems and Service Management*, Changdu, China, Jun. 2007, pp. 1–4. doi: 10.1109/ICSSSM.2007.4280163.
[8]. "Machine Learning Random Forest Algorithm - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed Aug. 28, 2021).
[9]. S. Khatri, A. Arora, and A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, Jan. 2020, pp. 680–683. doi: 10.1109/Confluence47617.2020.9057851.
[10]. "Techniques.pdf."
[11]. "Dealing with Imbalanced dataset. Techniques to handle imbalanced data | by Vaibhav Jayaswal | Towards Data Science." https://towardsdatascience.com/dealing-with-imbalanced-dataset-642a5f6ee297 (accessed Aug. 28, 2021).
[12]. S. Rajora *et al.*, "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, Nov. 2018, pp. 1958–1963. doi: 10.1109/SSCI.2018.8628930.
[13]. N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, p. 12, 2020.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details