iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43587

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Machine Learning Approach for Credit Card fraud Detection

Shruti Mohan Ghatge[1], Javeriya Shamshodin Shaikh[2], Arati Anil Kadam[3], Huzefa Riyaz Sayyed[4], Simran Sadique Shaikh[5], Prof. Ms. A.K Salunke[6]

[1, 2, 3, 4, 5]*Dept.of Computer Science and Engineering, Brahmdevdada Mane Institute Of Technology, Solapur*

[6]*Prof. in Dept.of Computer Science and Engineering, Brahmdevdada Mane Institute Of Technology, Solapur*

Abstract: Finance fraud is a growing problem with far consequences in the financial industry and while many techniques have been discovered. Data analysis is to be applied to finance databases to automate analysis of huge volumes of complex data. Data analysis has also played a salient role in the detection of credit card fraud in online transactions. Fraud detection in credit card is a data analysis problem, It becomes challenging due to two major reasons–first, the profiles of normal and fraudulent behaviors change frequently and secondly due to reason that credit card fraud data sets are highly skewed. This project propose investigation and to check the performance of various algorithms on highly skewed credit card fraud data. Dataset of credit card transactions is sourced from European cardholders containing 284,786 transactions. These techniques are applied on the raw and preprocessed data. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, precision.

I. INTRODUCTION

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions Become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set analysis,

Machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree and random forest is carried out Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behavior. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the Results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes a limitation to exchange of ideas and methods in fraud detection, and especially in Credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviors always different that is the legit transaction in the past may be a fraud in present or vice versa. This paper evaluates four advanced data analysis approaches, Decision tree, support vector machines, Logistic regression and random forests and then a collative comparison is made to evaluate that which model performed best. Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature (variables) selection for the models, suitable metric is most important part of data analysis to evaluate performance of techniques on skewed credit card fraud data. A number of challenges are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legitimate ones, Credit card fraud detection performance is greatly affected by type of sampling approaches to be use, selection of variables and detection technique to be use.

II. 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. This model is then used to identify whether a new transaction is fraudulent or not. Our aim here is to detect the fraudulent transactions.

III. OVERALL SCENARIO

A key challenge with payments fraud data is class imbalance. In the Kaggle dataset, roughly 99.8 percent of the transactions are labeled as legitimate and 0.2 percent as fraudulent. Class imbalance can make it difficult for standard models to learn to distinguish between the majority and minority classes [3]. As part of this project, I explore and assess methods to address this issue, including sampling techniques such as under sampling the majority class and oversampling the minority class. Due to privacy concerns, it is not possible to infer meaningful relationships between most of the variables in the dataset. For this reason, the focus of my project is on predictive performance rather than inference.

IV. REQUIREMENT SPECIFICATION

There is a long history of using machine learning for fraud detection in the payments industry. Bhattacharyya et. al. note that while fraud algorithms are actively used by banks and payment companies, the breadth of studies on the use of machine learning techniques for payment fraud detection is limited [4], possibly due to the sensitive nature of the data. Their study concluded that random forests, though not widely deployed, may outperform more traditional methods.

V. LITERATURE REVIEW

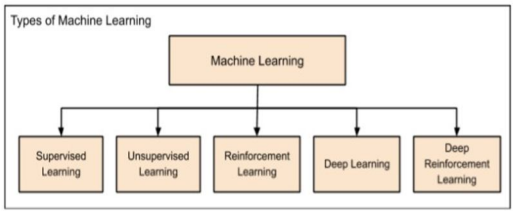
Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data analysis etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Analysis on Credit Card Fraud Detection Methods is to be done. In this project the survey is to be purely based on detecting the efficiency and transparency of each method. Significance of this project is to conduct a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem. And in the end conclusions about results of models' evaluative testing is to be made.

VI. COMPARING ALGORITHMS

In In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to 23% is obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential. Accuracy of an algorithm is around 50%. In Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system

VII. MACHINE LEARNING – CONCEPTS

Python has libraries that enables developers to use optimized algorithms. It implements popular machine learning techniques such as recommendation, classification, and clustering. Therefore, it is necessary to have a brief introduction to machine learning before we move further.



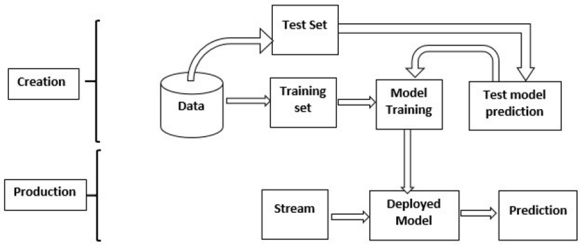
1) *Supervised Learning:* Supervised learning is analogous to training a child to walk. You will hold the child's hand, show him how to take his foot forward, walk yourself for a demonstration and so on, until the child learns to walk on his own.
2) *Unsupervised Learning:* In unsupervised learning, we do not specify a target variable to the machine, rather we ask machine "What can you tell me about X?". More specifically, we may ask questions such as given a huge data set X, "What are the five best groups we can make out of X?" or "What features occur together most frequently in X?". To arrive at the answers to such questions, you can understand that the number of data points that the machine would require to deduce a strategy would be very large. In case of supervised learning, the machine can be trained with even about few thousands of data points. However, in case of unsupervised learning, the number of data points that is reasonably accepted for learning starts in a few millions.

3) *Reinforcement Learning:* Consider training a pet dog, we train our pet to bring a ball to us. We throw the ball at a certain distance and ask the dog to fetch it back to us. Every time the dog does this right, we reward the dog. Slowly, the dog learns that doing the job rightly gives him a reward and then the dog starts doing the job right way every time in future. Exactly, this concept is applied in

"Reinforcement" type of learning.

VIII. METHODOLOGY

Ability of system to automatically learn and improve from experience without being explicitly programmed is called Machine Learning and it focuses on the development of computer programs that can access data and use it learn for themselves. And classifier can be stated as an algorithm that is used to implement classification especially in concrete implementation, it also refers to a mathematical function implemented by algorithm that will map input data into category. It is an instance of supervised learning i.e. where training set of correctly identified observations is available. The data set is splited the data into Test set, Train set.



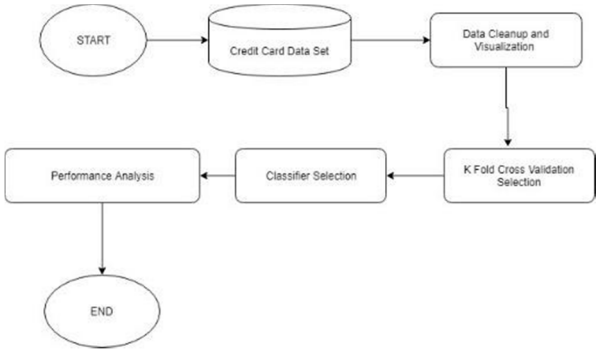
IX. PYTHON MACHINE LEARNING – ENVIRONMENT SETUP :

Libraries and Packages to understand machine learning, you need to have basic knowledge of Python programming. In addition, there are a number of libraries and packages generally used in performing various machine learning tasks as listed below:

1) *Numpy:* Is used for its Ndimensional array objects.
2) *Pandas:* Is a data analysis library that includes dataframes
3) *Matplotlib:* Is 2D plotting library for creating graphs and plots.
4) *Scikit-learn:* The algorithms used for data analysis and data mining tasks. **Seaborn** – a data visualization library based on matplotlib.

X. TRAINING THE DATASET

1) The first line imports iris data set which is already predefined in sklearn module. Iris data set is basically a table which contains information about various varieties of iris flowers.
2) We import kNeighborsClassifier algorithm and train_test_split class from sklearn and numpy module for use in this program.
3) Then we encapsulate load_iris() method in iris_dataset variable. Further we divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values (eg. petal length etc) and y prefix denotes target values (eg. 0 for setosa, 1 for virginica and 2 for versicolor).
4) This method divides dataset into training and test data randomly in ratio of 75:25. Then we encapsulate KNeighborsClassifier method in kn variable while keeping value of k=1.



This method contains K Nearest Neighbor algorithm in it.

5) In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

XI. TESTING THE DATASET

1) Now we have dimensions of a new flower in a numpy array called x_new and we want to predict the species of this flower. We do this using the predict method which takes this array as input and spits out predicted target value as output.
2) So the predicted target value comes out to be 0 which stands for setosa. So this flower has good chances to be of setosa species.
3) Finally we find the test score which is the ratio of no. of predictions found correct and total predictions made

XII. WORK CARRIED OUT

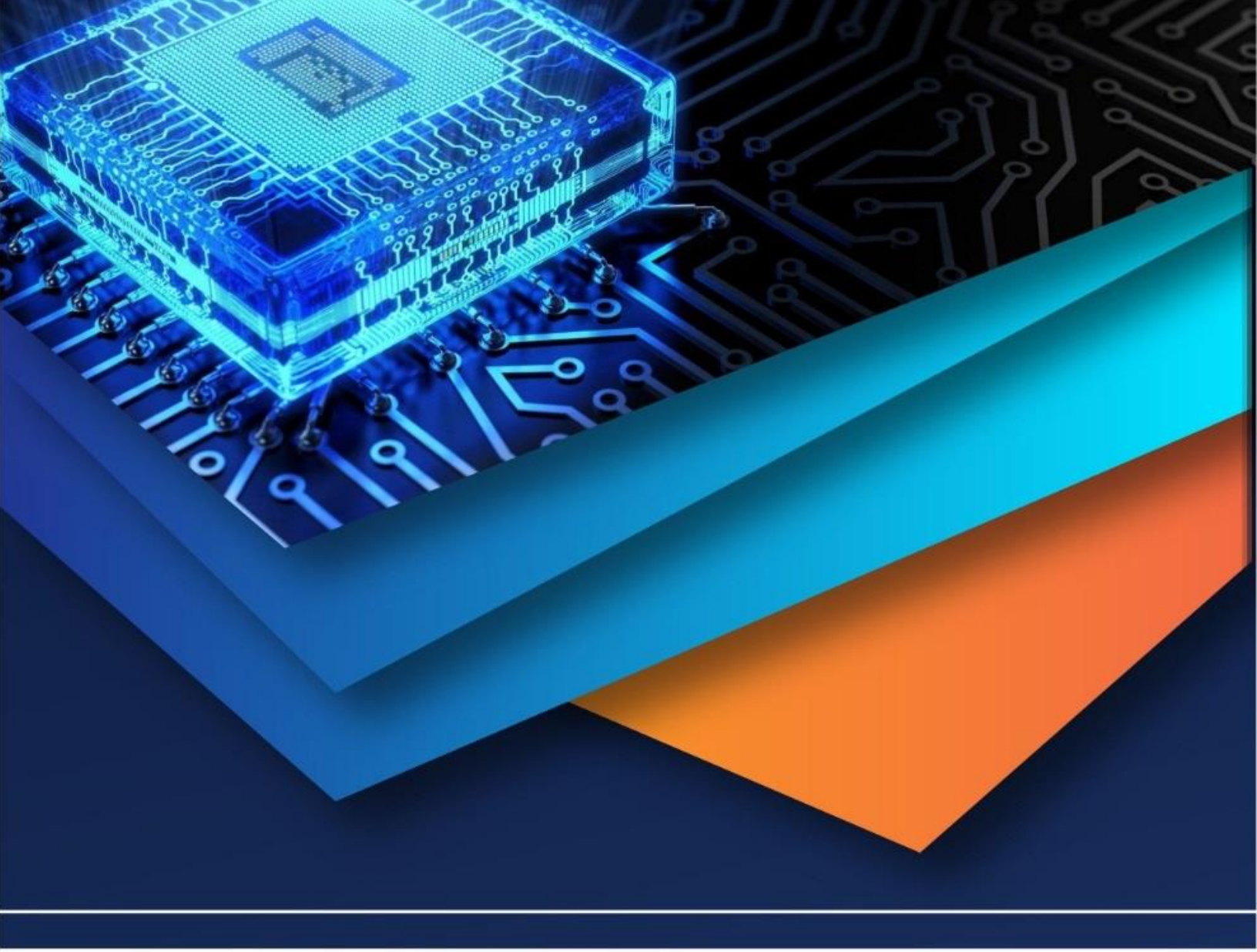
While we couldn't reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here.
The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result.
This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.
More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.
The same machine learning model of Credit Card fraud detection can be implemented in Run time.

XIII. CONCLUSION

From this project the result that is to be observed that different algorithms gives variation in accuracy, As Random Forest has a high accuracy in fraudulent detection while other shows medium accuracy from past experiments. So here we propose to check the accuracy of the prediction with various classifiers .And to deal with the big data set we opt using Big Data Analysis.

REFERENCES

[1] Dermala N., Agrawal A.N., Credit card fraud detection using SVM And Reduction of false alarms, International Journal of Innovations in Engineering And Technology (IJIET) 7(2) (2016).
[2] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Costsensitive credit card fraud detection using Bayes minimum risk.12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.
[3] Hafiz K.T., Aghili S., Zavarsky P., The use of predictive analytics Technology to detect credit card fraud in Canada, 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016), 1-6.
[4] Sonepat H.C.E., Bansal M., Survey Paper on Credit Card Fraud Detection, International Journal of Advanced Research in Computer Engineering & Technology 3(3) (2014).
[5] Worldline and the Machine Learning Group. Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/mlgulb/creditcardfraud.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY