

Fraud Detection using Machine Learning Approaches

Venkatachalam Vardhamana¹, B.Samatha²

^{1,2}Department of Computer Science and Systems Engineering
Andhra University College of Engineering, Visakhapatnam, AP, India

Abstract: In today's world, people heavily depend on financial institutions to take Credit Cards for Utilizing Banking Services like ATM Services, Online Transactions, Funds Transfer etc. Making use of this Services has become a part of daily life. The rapid increase of online transactions causes the Financial Institutions Cannot Guarantee to differentiate with fraudulent transactions with legitimate transactions. Due to This Financial institution Suffering Huge Loss of Money and customer loyalty is declining due to fraudulent transactions. The number of transactions in the banking sector is rapidly increasing and huge data volumes are available which represent the customer's behavior and the risks around the Fraudulent transactions. So Financial Institutions and Credit Card Issuers always need a more accurate predictive modeling system for many issues. Predicting credit Card Fraud Detection is a crucial task for the banking industry. Machine learning is one of the promising Context to extract patterns from the huge volumes of the data. This paper using different machine learning algorithms to build an efficient model to improve the accuracy of Credit Card Fraud Detection.

Keywords: -Machine Learning, Sampling, Pyspark, Logistic regression, Gradient Boosting Tree, Random Forest, Hadoop Distributed File System (HDFS).

I. INTRODUCTION

Financial Fraud is Alarming Problem in Almost Every Organization like Government, Corporate Organizations and Finance industry. Fraud can be defined as a

deliberate deception to secure unfair or unlawful gain or to deprive a victim of a legal right. credit card Based Payments Most Used in Now a Days, so there is Also Risk in Credit Card Transactions Occurred mainly in inner card fraud or external card fraud. The inner card fraud occurred due to consent between bank and cardholders, external fraud may Be Defined as Misleading usage of Stolen Credit cards. if you Detect Fraudulent transactions using Traditional Methods it will take more time meanwhile fraudulent Transactions may increase more. So Using Traditional Methods More Time Consuming and Inefficient. In today's Most of the financial institutions used Latest Computation Methodologies to handle Certain Credit Card Frauds. Fraud detection Take place when to Differentiate from fraudulent transactions with legitimate transactions. Transactions are categorized into two Classes of legitimate and fraudulent transactions. This Analysis based on Cardholder spending Behavior in certain Timing, Monthly Usage, Frequent Transactions Etc. Machine Learning algorithms will be used to study the historical credit data to extract patterns from it, which would help in predicting the likely Credit card fraud detection, thereby helping the financial institutions for making better decisions in the near future.

II. LITERATURE REVIEW

Suraj Patel, Varsha Nemade and Piyush Kumar Soni discussed a Big data analytical framework to process a large volume of data and to implement various machine learning algorithms for fraud detection. Observed their

performance on benchmark dataset to detect frauds on a real-time basis thereby giving low risk and high customer satisfaction to improve the analytical accuracy of fraud prediction, they have implemented three different analytical techniques. These analytical models are run on credit card dataset and accuracy of the analytical model is evaluated with help of confusion matrix. Among the three models, the random forest decision tree performs best in terms of accuracy, precision, and recall [1]. Masoumeh Zareapoor, Pourya Shamsolmoaliab trained various data mining techniques for credit card fraud detection. After several trial and comparisons; they introduced the bagging classifier based on decision tree, as the best classifier to construct the fraud detection model. The performance evaluation is performed on real-life credit card transactions dataset to demonstrate the benefit of the bagging ensemble algorithm [2]. Masoumeh Zareapoor, Afshar Alam, Seeja K.R presented a survey of various techniques used in credit card fraud detection and evaluates each methodology based on certain design criteria. And this survey enables us to build a hybrid approach for developing some effective algorithms which can perform well for the classification problem with variable misclassification costs and with higher accuracy [3]. John Richard D. Kho and Larry A. Veá suggested 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 and the Random Tree and J48 yielded the highest accuracy value of 94.32% and 93.50% respectively. By thorough analysis of these two (2) classifiers, it shows that the J48 is more fit in understanding the transaction logs data [4]. A hybrid technique of undersampling and oversampling is carried out on the European Credit card transaction dataset which contains 284807 records. Three techniques naive Bayes, k-nearest neighbor, and logistic regression classifiers are implemented and performance of the techniques is evaluated based on accuracy, sensitivity, specificity, precision, Matthews correlation coefficient,

and balanced classification rate [5]. Anusorn Charleonnán proposed the fraud detection of credit card payment by using a machine learning technique called RUSMRN. Proposed method adopts three base classifiers which are MLP, NB, and naive Bayes algorithms. This research is focusing on the information of the credit card company of Taiwan for collecting data on customer behaviors in credit card payments[6]. German E. Melo-Acosta, Freedy Duitama-Munoz, and Arias-Londono proposed a methodology for automatic detection of fraudulent transactions. This Methodology is based on a Balanced Random Forest, that can be used in supervised and semi-supervised scenarios through a co-training approach [7] [11]. Tina R. Patil and Mrs. S. S. Shereka put a light on performance evaluation based on the correct and incorrect instances of data classification using Naïve Bayes and J48 classification algorithm. The paper concluded that the accuracy of j48 is better than that of Naïve Bayes [9]. To stop the fallacious transactions a technique is designed which uses the combination of Hidden Markov Model, Behavior Based Technique, and Genetic Algorithm. Each and every transaction is tested with the above-mentioned technique [8]. Joseph Pun and Yuri Lawryshyn studied over 1 million unique credit card transactions from 11 months of data from a large Canadian bank. A meta-classifier model was applied to the transactions after being analyzed by the Bank's existing neural network based fraud detection algorithm. This meta-classifier model consists of 3 base classifiers constructed using the decision tree, naïve Bayesian, and k-nearest neighbor algorithms. The naïve Bayesian algorithm was also used as the meta-level algorithm to combine the base classifier predictions to produce the final classifier. Results from the research show that when a meta-classifier was deployed in series with the Bank's existing fraud detection algorithm improvements of up to 28% to their existing system can be achieved [10]. V. Mareeswari; G. Gunasekaran proposed a hybrid approach using support vector machine (HSVM) along with communal and spike detection for credit card application fraud detection [12].

III. PROPOSED METHODOLOGY

Data collection: The proposed system implementing algorithms on the historical credit card data which is collected from the ULB Machine Learning Group repository. The dataset contains 31 columns of 2,84,807 records, which describes different features of the dataset.

Data Preprocessing: data preprocessing is an important task to be done prior to analysis to get the data ready for analysis. As good data can only provide better results, data preprocessing becomes necessary prior to analysis. In data preprocessing, the proposed system performs, data imputation, data normalization, and transformation. Because of Real-world data is often incomplete or inconsistent, and it has to lack in certain behaviors or trends and is likely to contain many errors in Dataset.

PySpark: Python is a powerful programming language for handling complex data analysis and data munging tasks. It has several in-built libraries and frameworks to do data mining tasks efficiently. However, no programming language alone can handle big data processing efficiency. There is always need for a distributed computing framework like Hadoop or Spark. Spark is a lighting fast data processing tool useful for both batch processing and stream processing. The proposed system also implemented the algorithms using Pyspark to improve the performance of the predictive models.

Hadoop Distributed File System (HDFS): The proposed system used the HDFS as the primary storage area for storing the historical credit card transaction data. It employs a Name Node and Data Node architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters. HDFS breaks the data into small chunks and distributes them to different nodes in a cluster, thus enabling highly efficient parallel processing. HDFS uses master and slave architecture. Each Hadoop cluster consists of a single Name Node that manages the file system operations and Data Nodes on individual compute nodes. The HDFS supports applications with

large datasets. Proposed system using Hadoop as a storage device to handle a huge amount of credit card fraud transaction data.

Sampling: To balance the imbalance dataset we are using the undersampling technique. Undersampling is one of the techniques used for handling class imbalance. In this technique, we undersample the majority class to match the minority class. So in our dataset random sample of non-fraud class to match a number of fraud samples. This makes sure that the training data has an equal amount of fraud and non-fraud samples.

Implementation: The Proposed model applies the sampling technique on the dataset to balance it. Machine Learning Algorithms are implemented on the processed data to compare the performance of the algorithms. A random forest tree is a supervised machine learning algorithm which is useful for both classification and regression. Different machine learning algorithms such as gradient boosting tree, random forest and logistic regression are implemented by the system to find the best predictive model. The proposed system used the spark engine to improve the accuracy of the classification model. This system also compared the accuracies of algorithms before and after feature selection and Pyspark based accuracies of the algorithms to select the best algorithm that predicts the credit card frauds effectively. The architecture of the proposed system is given in fig.1.

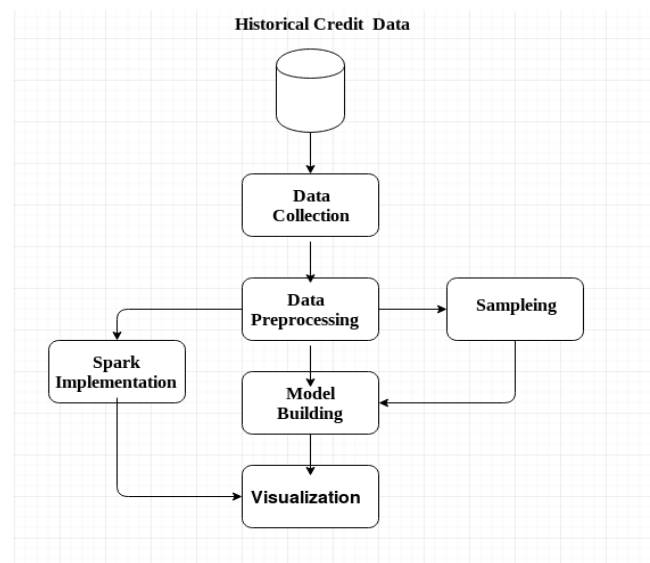


Fig.1 system architecture

IV. RESULTS

Machine Learning algorithms are trained and tested on the preprocessed data to find the model which predicts the credit card frauds with more accuracy. The accuracies of the model are shown in below tables.

Table.1 Performance of algorithms before sampling

Model	Accuracy	Precision	Recall
Logical Regression	99.92	88.34	61.90
Random Forest	99.94	93.91	73.46
GBT	99.25	83.2	70.74

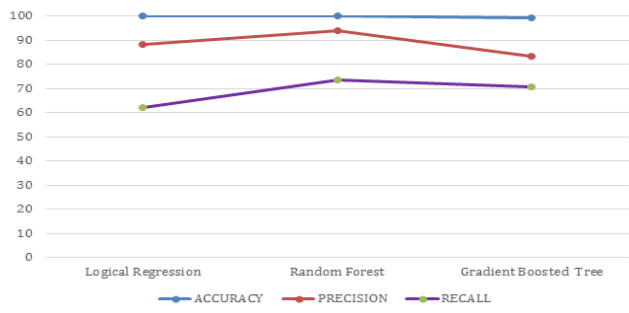


Fig.2 Before Sampling Comparison between Algorithms

Table 1 and Fig 1 showing the accuracies of the algorithms before sampling. The result clearly indicating that recall value for the models is low compared with precision values this is due to an imbalance of data in the dataset. The roc curves for the algorithms are shown in below figures.

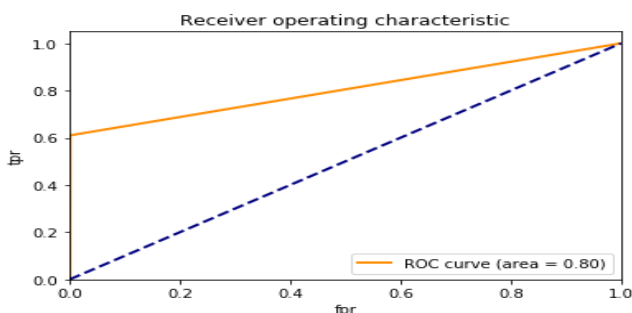


Fig.2. Logical Regression ROC Curve

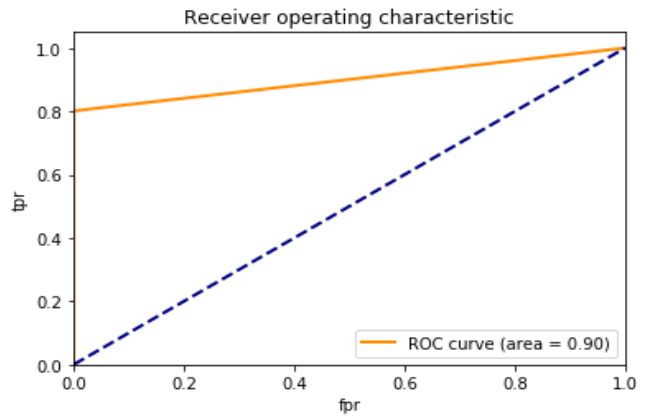


Fig.3 Random Forest ROC Curve

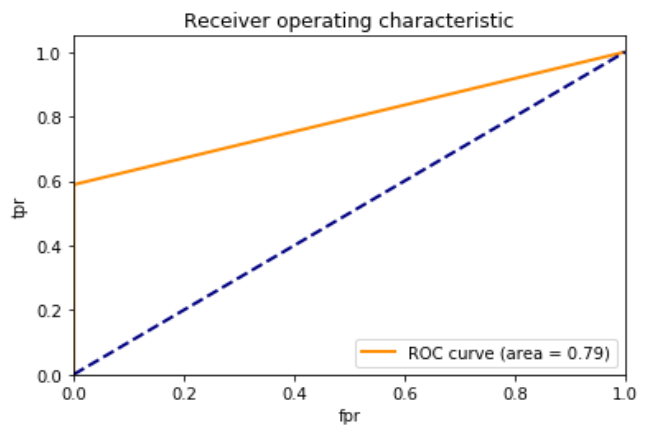


Fig.4. GBT ROC Curve

Table.2 Metrics of algorithms after sampling

Model	Accuracy	Precision	Recall
Logical Regression	94.25	95.77	92.31
Random Forest	95.93	97.14	93.51
GBT	93.91	95.10	92.41

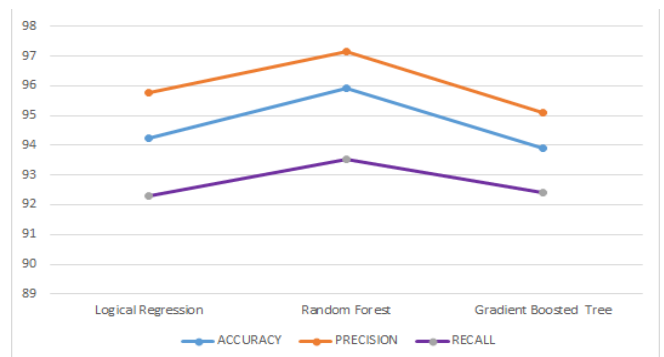


Fig.5 After Sampling Comparison between Algorithms

Table 2 and Fig.5 indicating the results of the algorithms after sampling technique. From the table, it is seen that the system has improved the recall values of algorithms after using the sampling technique. roc curves are shown in below figures.

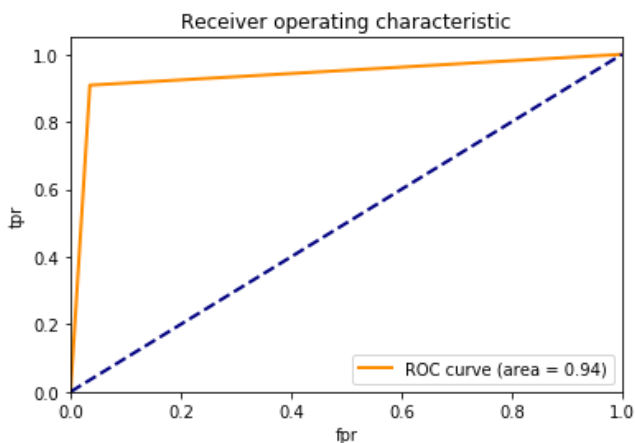


Fig.6 Logical Regression ROC Curve

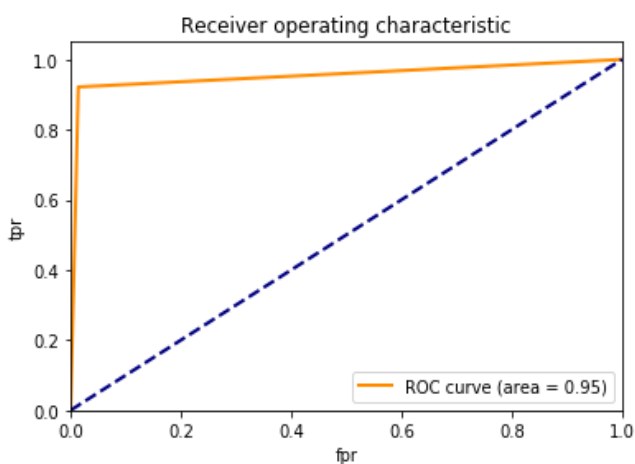


Fig.7 Random Forest ROC Curve

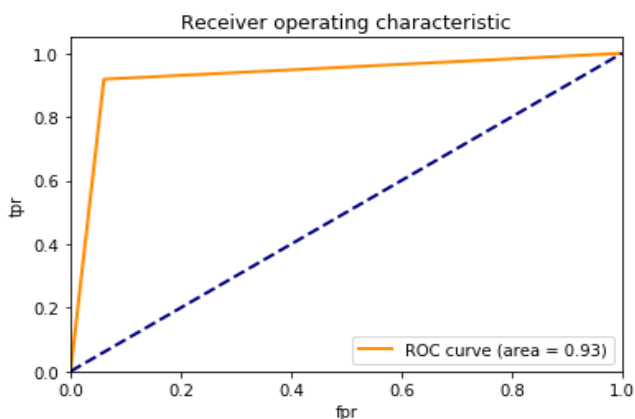


Fig.8. GBT ROC Curve

Table.3 Performance of the algorithms using Pyspark

Model	Accuracy	Roc
Logical Regression	98.90	97.02
Random Forest	99.93	98.31
GBT	98.93	97.49

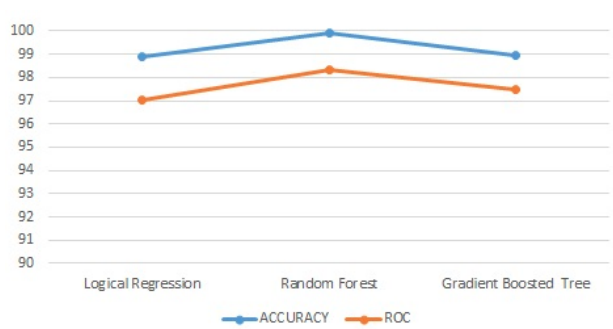


Fig.9 Comparison of Algorithms using PySpark

it could be observed that our proposed Random Forest recorded better performance with Accuracy of 99.93 using Pyspark.

V. CONCLUSIONS

In this paper, in order to predict credit card fraud detection, historical credit data collected from the UCI repository. The collected data is not in balanced mode. So to balance the dataset, a sampling technique is applied to it. In this work, we used different Machine Learning algorithms for credit card fraud detection. The study of the results indicating that the Random Forest algorithm outperforming the other models.

REFERENCES

[1] Suraj Patel, Varsha Nemade, Piyush Kumar Soni, "Predictive Modelling for Credit Card Fraud Detection Using Data Analytics", in International Conference on Computational Intelligence and Data Science, 2015.

[2] Masoumeh Zareapoor, Pourya Shamsolmoaliab, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier," in International Conference on Intelligent Computing, Communication, and Convergence, 2015.

- [3] Masoumeh Zareapoor, Afshar Alam, Seeja K.R, “Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria”, in International Journal of Computer Applications, 2012.
- [4] John Richard D. Kho, Larry A. Veal, “Credit card fraud detection based on transaction behavior”, in Tencon IEEE Region 10 Conference, 2016.
- [5] John O. Awoyemi, Adebayo O. Adetunmbi, “Credit card fraud detection using Machine Learning Techniques,” 2017.
- [6] Anusom Charleonnann, “Credit Card Fraud Detection Using RUS and MRN Algorithms”, in International Conference on Management and Innovation Technology, 2016.
- [7] German E. Melo-Acosta, Freedy Duitama-Munoz, Arias-Londono, “Fraud Detection in Big Data using Supervised and Semi-Supervised Learning Techniques,” 2017.
- [8] Ayushi Agrawal, Shiv Kumar, Amit Kumar Mishra, “A Novel Approach for Credit Card Fraud Detection”, in International Conference on Computing for Sustainable Global Development, 2015.
- [9] Tina R. Patil, Mrs. S. S. Sherekar, “performance analysis of naive Bayes and j48 classification algorithm for data classification”, in International Journal of Computer science and applications, 2013.
- [10] Joseph Pun, Yuri Lawryshyn, “Improving credit card fraud detection using a meta-classification strategy,” 2012.
- [11] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang, “Random forest for credit card fraud detection,” 2018.
- [12] Mareeswari; G. Gunasekaran, “Prevention of credit card fraud detection based on HSVM,” 2016.