Financial Fraud Detection Using Machine Learning

By

Under the guidance of
Chandan Mishra

Table of content

Abstract:

Due to boom of smart phones electronic money transfer through a device or mobile has grown drastically in the last few years. The ease of receiving and transferring money has not only attracted the users but this has caught attention of fraudsters and attackers as well. The traditional rules applied methods take long time and labor for fraud detection techniques. As the amount of data is increasing widely, the traditional rule-based method wont be able to deal with such heavy data on time. Therefore flexible computational methods are used such as data mining and machine learning in detecting fraud as they are very critical methods to identify the illegal fraud activities by trying to rule out all possible false alarms and increase true positive frauds. In our research we debate on the financial fraud detection problem as binary classification. Machine learning techniques have been applied to detect whether a transaction is valid or fraud. Our study has been applied using five machine learning classification algorithms namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K Nearest Neighbor (KNN) and XGBoost (XGB). These models are good fitting and good ability to learn from the accuracy as it displays deep insights and outcome of the models used. From the five models three models perform exceptional and equally good based on their accuracy, confusion matrix, ROC and classification reports. However the accuracy of these three models given as XGB 93.2%, RF 92.9% and DT 92.8%. From these models we can see that XGB is the best model on practical results but RF is the superior model based on the cost of models. But we can conclude that all three models are best suitable for the prediction of fraud for this specific dataset and machine learning tools are most important in solving these problems.

List of Figures

1.Introduction

Different types of fraud detections have boundless variations and Financial Fraud Detection is been in the globe since ages. Financial transactions are executed with an ease due to quick growth of banking sectors and this has availed fast and smooth communication, hereinafter there is immense expansion in the count of frauds causing to the loss of billions of dollars. Past few decades, government, international organizations and financial institutions have done a few lateral laws, regulations and applied various advance approach to analyze and avoid such kind of fraudulent exercises.

The definition of Fraud given by "Association of Certified Fraud Examiners (ACFE) The ACFE Association of Fraud Examiners Certificates, fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception or other unfair acts." (V, M, K, M, & R, 2017) Various sorts of financial fraud are mentioned below:

1. Credit Card Fraud
2. Financial Statement Fraud
3. Mortgage Fraud
4. Securities and commodities fraud.
5. Money Laundering

The recent few years a type of electronic money transfer which is mobile money transaction having no involvement of any bank accounts is enormously expanding fast since advancement of smartphones and online services. Banking relationships and financial facilities are distinguished by mobile money transaction hence it is largely spread in developing countries. Mobile money transactions systems require financial institutions with subject to same controls, illegal mobile money transactions which include other activities like money laundering etc. Mobile money transactions is famous for leading fraudulent activities and attacks. Thus there are numerous smart computing technologies from whom data mining and machine learning are the utmost crucial solutions that help in all these illegal activities detection.

There is lack of public data in this sort of topic to have a relative comparison based on different methods which is the limitation of the research. In many papers authors have consumed their personal data which has confidential and privacy issue thus disclosure of it is void for the vulnerabilities of services and company clients privacy to be maintained on accountability. Thus for the purpose of research PaySim dataset is used which has simulated dataset containing real and imitated data.

The paper consist of classifications machine learning algorithms, different algorithms are applied for fraud detection research such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K Nearest Neighbor (KNN) and XGBoost (XGB). There is sufficient training data which consist of non-fraud transactions and fraud transactions as well, the object is to affix machine learning algorithm to train the historic data so that it can predict if the transactions taking place are

non-fraud or fraud. The XGBoost model outperforms the results from the algorithms applied for comparison with other models comparing to the mentioned papers. Experiments are conducted on the dataset and after balancing data the results show that XGBoost surpass the rest state-of-art models on the basis for accuracy, confusion matric and classification report.

Research Question:

Comparison of machine learning techniques for prediction of financial fraud detection.

2.Literature Review

From the various forms of fraud, Financial fraud detection has grown tremendously and requires lot of consideration and involvement. Financial frauds upset both the financial corporation and day to day living(Sadagali, Sael, & Benabbou, 2019,3). This lowers the industries courage, it diminishes savings and cost of living increases. Various prevention models are used by Financial institutes to detect fraud and consign the issue. There are numerous ways of intruding all defensive models perceive that the fraudsters are adaptive. Even after great attempt the government, law enforcement and financial institutions the fraud keeps on exponentially increasing. Nowadays the fraudsters would be fast fraternity, creative and very innovative.
Our research investigate and impose comparative analysis of financial fraud detection techniques, such as machine learning techniques, which perform a crucial act in detection of financial fraud, it usually unveils and exposes the mysterious truth of huge volume of data. Lots of modernized techniques are continuously involved for fraud detection and are used in different areas because of exceptional removal of fraud which affects financial category every year. Financial fraud detection will focus on its weaknesses and describe the problems of fraud analysis. Hence the literature review shows different researchers and the various techniques applied on the topic.

(Dahee & Kyungho, 2017), proposed the approach where the main process of fraud detection for mobile payment is based on unsupervised, supervised and machine learning methods which can detect big volume of financial fraud data. In the paper feature engineering and sample processing is applied for quick processing of huge amount of transactions and to gain better accuracy in mobile payment fraud detection. The research of the paper have evaluated results by implementing Fmeasure and ROC curve. The paper has both supervised and unsupervised methods which is different than most familiar observed single context of data mining method. Approach of semi-supervised method was able to label new types of fraud and has been most effective method for mobile payment fraud transactions. The semi-supervised approach was straight classification effective in this type of fraud detection and better precision under regression. The assessment is established by F1 result and AUC.

(Jianrong, Jie, & Lu, 2018), paper displays combination of machine learning classification models and feature selection for optimization of financial fraud detection. The document has proven three points. One, it analyzes and calculates the factors which effect the behavior of fraud. The debate states that random forest is the outstanding model among the four applied models. Two, the model is influence by number of variables. Output lists that either 2 or 5 variables give better results than others. Three, researches applied five machine learning models and the performance was compared and it is found that between these five models in two feature selection Xgboost performed good but random forest has edge points over the rest: 1. High dimension data processing is better. 2. It can neglect overfitting to some limit. 3. Model has balanced result and sound robustness.

Drawback is that the research data is not big adequately and the variable numbers can be more flexible. Practically there are many complexion for factors and variable design.

(Prabin, 2011), the framework used in the paper is known as
"(KDIFD)Knowledge-driven Internal Fraud Detection" which is planned to detect internal financial frauds. The process is a systematic approach where auditors can explore internal financial frauds. It helps auditors in verifying the possible fraudulent or non fraudulent activity based on appropriate data, context, preparing the data and sourcing in chosen data structure, transferring and cleaning into responsive form, picking the technique, data mining and analyzing. The research consist of both machine based data analysis and auditor's knowledge base and data mining techniques. The author describes further research for real time fraud detection.

(Adeyinka, Stelios, Miltos, & Emmanouil, 2016), the authors state a Multi intelligent Fraud Detection System applying Neural Network (NN), Logistic Regression (LR) and Case base reasoning (CBR). The performance evaluation is done on a simulated data which is synthetic used to prove the efficiency of the methods. On comparing the models Logistic Regression classifier has good outcome among the three methods, it has substantial growth in sensitivity, precision and specificity as the ratio for train and test was changed. Under analysis the data used is described with incomplete information, uncertainty, linearity and fuzziness which is because of corporate, legal, societal and sensitivity impact of public domain information which is confidential. This makes fraud detection difficult as exploring in depth data has issues specifically with concentrating on monitoring and tracking transaction array with characteristics and similarities in different categories of fraud handling contained procedure. This is why CBR is implemented. In future the authors decide to program the model with bit if uncertainty and fuzziness to develop knowledge pool of various fraud patterns and implement CBR for the case base.

(G. et al., 2009), exhibited generic financial fraud detection framework to classify and understand various combination of data mining and financial fraud detection techniques. Based on different evaluation criteria the framework permits us to determine features variety of financial fraud detection algorithms. According to paper the future work depends on future work as there is no consent that which features in data are good for detecting fraud. It is required to merge financial data to more information like proportion, governance style and auditor size for final analysis. However advance technologies will have more specific financial fraud detection methodologies and have wider applications where combination of multi type data will take place. The paper discusses about ensemble unsupervised and supervised methods which will give better outcome in future.

(Jarrod, Maumita, & Rafiqul), proposed intelligent computational and statistical concept on the ongoing practices in the financial fraud detection. The approach varied the classification performance based on accuracy, sensitivity and specificity of all the techniques and was displayed

to be consistent and capable of detecting different financial fraud. Competence of computational intelligent (CI) methods like support vector machines and neural networks to understand and acclimate to new situation is extremely effective at overpower the emerging tactics of fraudsters.
Further future work is the variations of all type of financial fraud which will drive thru a generic framework and enormously build up the area of intelligent financial fraud detection for the domain.
(Mahdi, Qingfei, Vahab, & Muhammad, 2019), this paper applies cluster analysis and segregates the data in three classes. Cluster analysis results are used in apply statistic experiments on research and test the recall values and precision of the few supervised methods such as Support vector machine(SVM), Probabilistic neural network (PNN), Multi nominal log-linear model(MLM), Multi-layer feed forward neural network (MFFNN),and Discriminant analysis (DA) for various conditions. Results of the paper show impact which is significant in most fraudulent cases, though it is not monotonic. Discriminant analysis (DA) and Multi-layer feed forward neural network (MFFNN) are the lone methods which are correlated by different proportion of fraud associations. Sample size affects the recall values and precision for all the methods, Larger the sample greater the precision for all methods. PNN has drop in trend for recall values except all other methods. Multilayer feed forward neural network (MFFNN) outperforms in all terms comparing the other methods. PNN still has greater value that MFFNN but the difference is minute and can be ignore for greater sample size.
The outcome of the paper can help auditors maintain cost and time by applying automated tools, and it saves work load effectively. Future work can be improved adding more including new variables will turn out to be definitive factor. The procedure and methodology in this paper can be used for other area for various purpose and not just for fraud detection.
(Sadagali et al., 2019), according to this study hybrid models are the most utilized fraud detection techniques, because they consist combination of various traditional models. The paper does not smother all categories of fraud and every category of fraud has specific hindrance, for the kind of response expected in real time and text analysis. The future work is based to improve ongoing algorithms on credit card fraud, authors want implement hybrid model which can handle real time problem and imbalanced dataset both so that it can respond to a financial transaction while runtime and with better accuracy.
(Ratha & Pech, 2019), The paper discusses binary classification which is supervised learning used for mobile money transfer in detection of financial fraud problem. They have included three machine learning models namely Naive Bayes (NB), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) for the solution. The models performed fairly on the PaySim simulated public dataset. According to author MLP has higher accuracy but NB is the best overall performing model. MLP has parameter which were altered to gain better results. The methods used being traditional perform well on the data. The paper states shortage of public data for scientific comparison and systematic approach. (CLIFTON, VINCENT, KATE, & ROSS), according to their research the paper covers all types of fraud detection study. The paper states all subtypes, types and adversary of fraud, technical aspects, performance metrics, all the techniques and methods. The paper states the area of research can be benefited with more related fields from determining

limitations in techniques and methods of fraud detection. For future work they want to apply unsupervised methods for study of counterterrorism which is text mining from law enforcement and system monitoring, and in semi-supervised method spam detection community and intrusion in game theoretic could help future of fraud detection improvement. Rasa Kanapickiene and Zivile Grundiene, 2015 [17] (Rasa & Zivile, 2015), paper consist of theoretical analysis where activity, profitability, liquidity and structure ratios are interpreted often. The experimental research explore 51 financial ratios. The values indicated in financial fraud detection is based on financial ratios. The author applied Logistic Regression (LR) for financial statements fraud detection distinguished by financial ratio. While taking decisions for company evaluation and investment the model can be utilized by exterior user for financial statement information.

(Suduan, 2016), proposed applying numerous data mining techniques by a non conventional analysis. The model included Decision Tree (DT), Bagging Boosting Network (BBN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) to build a better accurate model for financial statement fraud detection. In the initial stage the paper applies DT of CART and CHAID model to choose decisive variables. Then they build a integrated classification model using CART, CHAID, SVM, ANN and BBN for comparative analysis. The result of model in paper is based on overall accuracy with CART and CHAID model is top performing with 87.97% and having least Type 1 error rate of 7.31. According to the experimental results of the paper accuracy of DT CHAID, with CART is higher in detecting financial statement fraud. The implementation can be used as a tool by auditors for financial statement fraud detection. The paper methods can contribute to shareholders, credit rating institutions, investors, company managers, security analyst, auditors, certified public accountants (CPAs), relevant academic institutions and financial regulatory authorities.

(QIAN, TONG, & WEI, 2009), have shown objective and subjective combination of methods for financial fraud detection systems. Rough set (RS) and AHP are deployed in the integrated framework to select intrinsic features, analyze the fraud possibilities, alarm and abnormities. Framework which is proposed focuses on identity theft for the financial fraud detection, and observes and avoid other categories of financial fraud systems. System developed is intelligent to apply on any open platform to different methods and algorithms. In future the paper states to design online intelligent fraud detection system with the proposed framework. The traditional methods and models will be compared with integrated method with real life or experimental data.

(Bian et al., 2016), paper expresses a concept for financial fraud detection classification for imbalanced data. The paper consist of certain contributions where they implement a framework that merges BI methods to ensemble technique with improved stability and accuracy for financial datasets having disarranged samples. Secondly BI literature and public financial fraud detection provide new insights to the understanding. The third shows that in this domain it is helpful to feature the role of BI analysis techniques. The advantage of bagging and boosting to decrease the variance will benefit investors, government regulators and audit firms. Paper future work the framework will be enhanced merging existing methods with other competing methods. The

proposed framework should facilitate BI techniques to execute better in terms of financial fraud detection.

(Prasad, Shuhao, & Yibing, IEEE Systems and Information Engineering Design Symposium), the paper proposes different natural language processing and supervised machine learning techniques such as Neural Networks, Latent Dirichlet Allocation (LDA), Binomial Logistic Regression, Support Vector Machines and Ensemble Techniques for the financial fraud detection. In paper LDA is applied on financial reports of type 10K and from reports set up frequency matrix and used the data for advance classification algorithms. The metrics authors evaluate are Area Under the Curve , Precision and Receiver Operating Characteristic Curve which give performance of all algorithms. The results of paper show text classifiers accuracy reaches levels as expected. The future work is to gain better accuracy by expanding data size and sum of topics created by LDA.

(Simon & Li, 2019), paper studies adversarial attacks which affect the fraud detection system based on a smartphone payment application. Paper compared different mobile payment detection for adversarial settings and compared them. Authors show adversarial attacks can happen on machine learning techniques of fraud detection systems. It debated training data balance and robustness of fraud detection models can be improved by adversarial examples. Adversarial and benign environment showed results where performance proposed by paper was improved in both environments. The future would be building a more simple framework for adversarial environment which can decide robustness of fraud detection systems.

(Yuh-Jen & Chun-Han, 2017), paper examines significance of big data in finance and economics and variety of characteristics to implement a big data financial statement fraud detection of corporate groups for precise detection of fraud within corporate groups and reducing risks and investment losses and improving investment decision which will profit creditors and investors.

Kunlin Yang, 2018 [25] (Kunlin, 2018), have exhibited Fraud Memory which is a novel fraud detection algorithm. The neural model is sequential and memory enhanced for financial fraud detection with strong interpretability. It influences memory based networks and captures sequential patterns to improve performance of both. The paper Complex operations are executed in this model suitably for financial systems. Accuracy depiction and objectives are done by logs and users. The model contains great performance with concept drift adaptability and lesser false positive rate. The model has flaws of cold start as detailed description is required for users which adds it to future work.

(Dongsong & Lina, 2004), this paper discusses about the data minig techniques which have been applied to trace the obscure pattern and future prediction trends with behaviors in the financial market. Data mining has achieved competitive advantages which consist reduced cost, increased revenue and much enhanced marketplace awareness and responsiveness. Lots of research communities are focusing on exploring the data mining techniques on solving the financial problems. The paper explains data mining with both perspective technical and application based on context of financial problems. Additionally the authors compare various data mining techniques and debate few important data mining issues involved in particular data mining techniques. The paper displays sum of challenges and trends for future research in this area. To achieve effective

financial management in both individual as well as institutions. Lately data mining techniques have presented better potentials in financial industry and will keep on blooming the fresh knowledge-based economy.

(Jianrong, Yanqin, Shuiqing, Yuangao, & Yixiao, 2019), The study of the papers depicts financial fraud activities dependent on 7 non financial and 17 financial variables by implementing few data mining techniques namely classification and regression tree (CART), support vector machine (SVM), logistic regression (LR), Bayes classifier (Bayes), back propagation neural network (BP-NN) and K-nearest neighbor (KNN). There were 134 companies from stock exchange of Shanghai and Shenzhen involved in fraudulent cases. The variable dimensionality was reduced by adopting principal component analysis (PCA) and ordered regression. The results with proposed models show that SVM has the best accuracy from all the data mining techniques applied and the accuracy of the classification was enhanced for all the implemented data mining techniques and 13 significant variables were screened by stepwise regression. The classification results were not superior for the initial 16 principal components transformed by PCA. Hence, SVM and stepwise regression dimensionality method together turned out to be better models for financial statements fraud detection.

3. Literature Review

Research methodology is special case of classification. It defines the present study and gives accurate analyses of the data. The classification used in the methodology is binary classification and consist of multi-class problems which just has data of two types of classes consumed in the modelling. The methodology displays details about different factors where the study is applied. The approach in the study is deductive based on the ability and analyses specific amount of data to display logical evaluation and edgy conclusions. One side of fraud detection could be described as binary classification in such a way where a transaction classification is either fraud or non-fraud (Ratha & Pech, 2019). The other side of fraud detection could be one class classification and outlier detection where the fraud transactions can be displayed as outlying transactions which explains them to be different from the regular transactions. This study consist of binary classification and we keep the other study for future work. The study is conducted with applying machine learning techniques where certain data is analyzed which is collected from some particular source which could be journals, books, websites or articles etc. The selected data will consist of numerical data and will be explored in the design of charts and graph. Therefore we can say that the analyses and data collection will be based on quantitative method. The brief of methodology would be using machine learning techniques and involving supervised learning algorithms. We choose five trendy machine learning algorithms named Logistic Regression (LR), Random Forest (RF), Decision Tree(DT), Extreme Gradient Boost (XGB) and K Nearest Neighbors (KNN). We discuss the following algorithms in details below.

3.1 Random Forest

Random Forest is build from decision tree. This is an ensemble method. The algorithm is supervised machine learning algorithm. The “bagging” method is used to train the random forest. In bagging method the overall result is improved by combining the learning models. This means random forest simply constructs numerous. (V et al., 2017)

Random forest is a flexible model as it can be used for regression as well as classification problems. The random forest classifications explained as we have used classification in our research. The image attached shows two trees and describes a random forest.

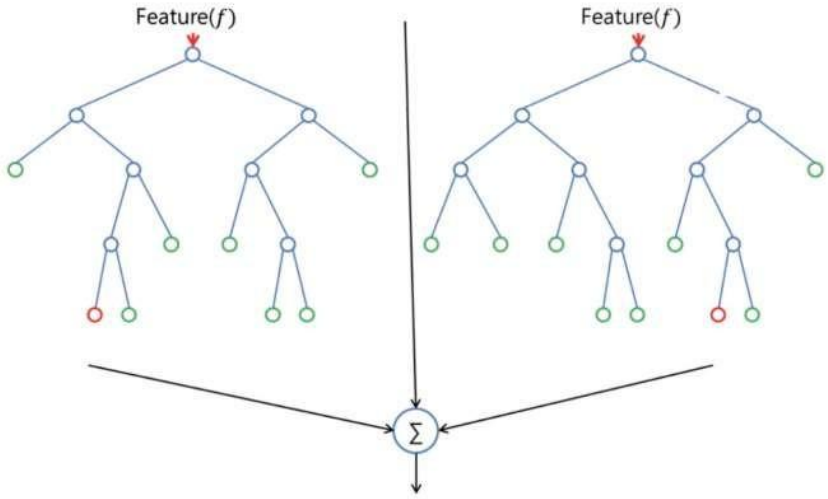

Figure 1. Illustration of Random Forest.

The hyper parameters in random forest are similar to bagging classifier or decision tree. Randomness is included in random forest by default when trees grow. The random subset of features are applied to get the best feature rather than searching for essential features during node splitting. The wide diversity results to a better model. Hence in random forest, applying the random thresholds to each feature instead of looking for finest thresholds.

The hyperparameters in random forest are either used to increase the predictive power of the model or to make the model faster. Let's look at the hyperparameters of sklearn built-in random forest function.

Hyperparameters of Random Forest.

There are two approaches in apply hyperparameters one is to improve the predictive power and second to boost model speed.

1. Improve Prediction Power.

   n_estimators : Decides number of trees to improve the performance and have stable predictions but n_estimators slower the computation of the algorithm as well.

   max_features : This selects max number of features to split a node.

min_simple_leaf : It splits internal node by determining minimum number of leaves required.

2. Boost Model Speed.

n_jobs : The hyperparameter decides how many processors show be used. When the value is one only one processor can be used and if value is minus one then no limit on using processors.

random_state : The values produced after defining a random state the model generates same results based on the applied training data and hyperparameters.

oob_score : The oob sampling is one third of the data not used to train the model and used for performance evaluation which is call out of bag samples.

3.1.1 Random Forest Approach

Random Forest classifier is used in the methodology to predict the fraud and to display if a transaction is a valid transaction of a fraud transaction. The machine learning algorithm is applied by selecting specific features which help to improve the prediction power of the algorithm. The improvement of the model is based on multiple decision trees and samples to see the dependency which smoother the prediction. The aim is to show whether a transaction is fraud or not by evaluating the random forest classifier with selecting certain hyperparameters which held in tuning the model for better results.

3.2 Logistic Regression

When the dependency of the variable is e.g. yes or no, true/false and not a number. It is called logistic regression which is a type of linear regression. This regression performs classification and classifies into classes based on dependent variable. It is hugely used in various domains and the modelling is done to build relationship between categorical outcome variable, like our topic where we have to predict whether a transaction is fraudulent or non fraudulent. (Jianrong et al., 2019)

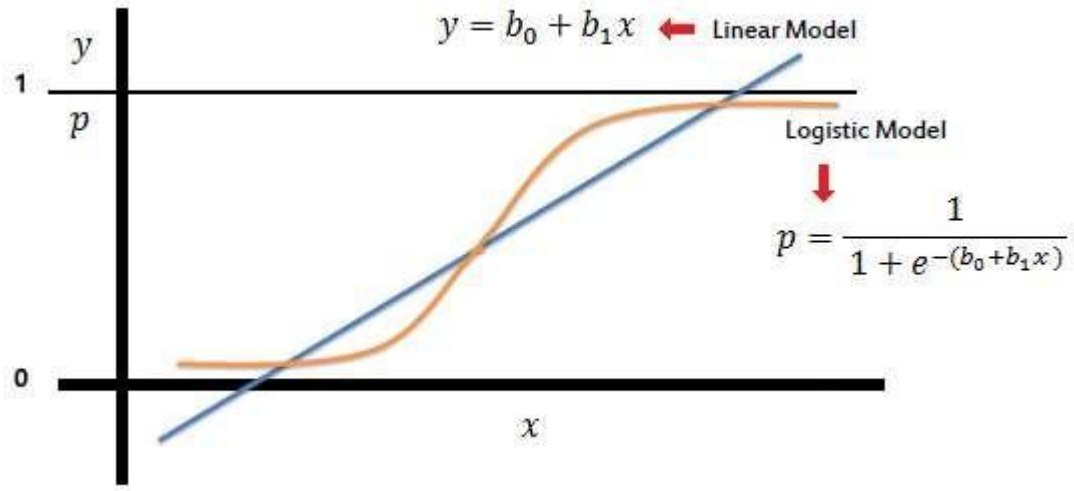

Figure 2. Illustration of Logistic Regression.

Binary output prediction is done in logistic regression. The model out put is based on two variable and it predicts which variable is applied more in the model.

$$y = b_0 + b_1 x$$

Linear Regression

The model linear regression is obtained by achieving relationship between variables. 0.5 is assumed as the threshold for the classification line.

$$p = \frac{1}{1 + e^{-y}}$$

Logistic Sigmoid Function

Probabilities of the class belong to regression or not is decided by applying logistic function.

The function is given as log of the probability of the occurring event to the log of the probability not occurring. Based on higher probability of one of the class the variable is classified.

3.2.1 Logistic Regression Approach

In Logistic Regression classification we focus on the capture most fraudulent transaction. As Logistic Regression is a binary classification algorithm it focus on true or false, 0 or 1. Where in our method we have target variable isFraud and the model predicts about the transaction that it is fraud or not. We predict the model to check the accuracy of the logistic regression.

3.3 Decision Tree

Decision tree is build from tree structures and it can be classification of regression models. Incremental development of an associated decision tree breaking down the dataset into tiny subsets and the resultants tree is formed with leaf nodes and decision nodes. The split is determined by Iterative dichotomiser 3 (ID3) algorithm structure.

A decision tree is implemented by entropy and information gain. (Sadgali, Sael, & Benabbou, 2019)

Entropy

Entropy is randomness of elements based on the uncertainty of degree or amount. Entropy is simply defined as impurity measure.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Entropy

Entropy gives predictability of particular events naturally. Homogeneity sample is calculated for entropy. The zero entropy sample is completely homogeneous. If the entropy is one it means sample is equally divided.

Information Gain

Information gain is based on independent attribute measures the relative change in entropy. Each attribute tries to estimate the information contained by them. The attribute with highest information gain helps in building of a decision tree i.e. maximum homogeneous branches.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Feature X is applied for information gain where Gain (T,X). The entire set of entropy is Entropy(T) and the other term is calculation based on the feature X which is Entropy (T,X).

A given node in the tree ranks attributes for filtering information gain. On each split the highest information gain entropy is ranked.

Overfitting is a disadvantage of decision trees, it reduces test accuracy when it tries to fit model by digging deep in train set.

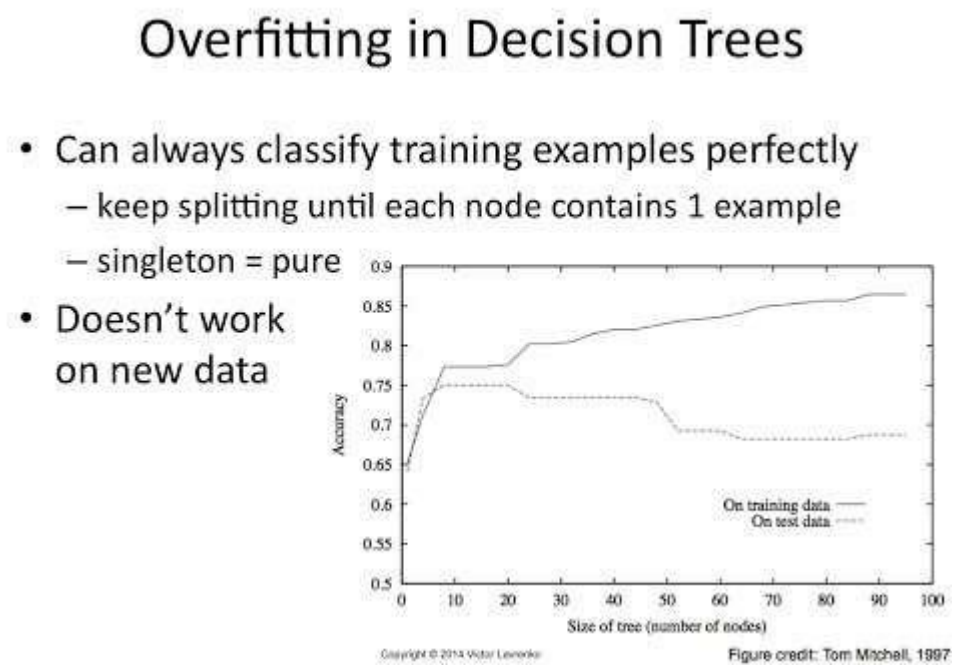

Figure 3. Overfitting Decision Trees

Pruning nodes can minimize overfitting in decision trees.

3.3.1 Decision Tree Approach

Decision Tree classifier is used in the methodology to predict the fraud and to display if a transaction is a valid transaction of a fraud transaction. The machine learning algorithm is applied by selecting specific features which help to improve the prediction power of the algorithm. The improvement of the model is based on multiple decision trees and samples to see the dependency which smoother the prediction. The aim is to show whether a transaction is fraud or not by evaluating the decision tree classifier with selecting certain hyperparameters which held in tuning the model for better results.

3.4 K-Nearest Neighbor

KNN is called lazy learning and non-parametric algorithm. It does not assume any data underlying in the distribution. Determination of the model structure is based on the dataset. Almost all worldwide dataset dont follow mathematical theoretical assumptions into practice. This algorithm model generation doesn't require any data points. For test phase all the training data is consumed. Due to this the testing phase is costlier and slower whereas training becomes faster. Memory and time increased cost in testing phase. KNN consumes more time in all data points scanning and this needs lots of storing data which utilizes lots of memory. In k nearest neighbor, number of nearest neighbor is K. The important decisive factor is number of neighbors. K is usually odd number if count of K is 2. If K=1 then it is said to be nearest neighbor model. (Jianrong et al., 2019)

3.4.1 KNN Approach:

Grouping of class labels are the essential approach for estimation of KNN. The main purpose of KNN is testing sample with category placed of given area with maximum region covered with K neighbors. In the figure below we can see that red class has covered more X region than the blue has covered in the selected area. So X area has red class classification.

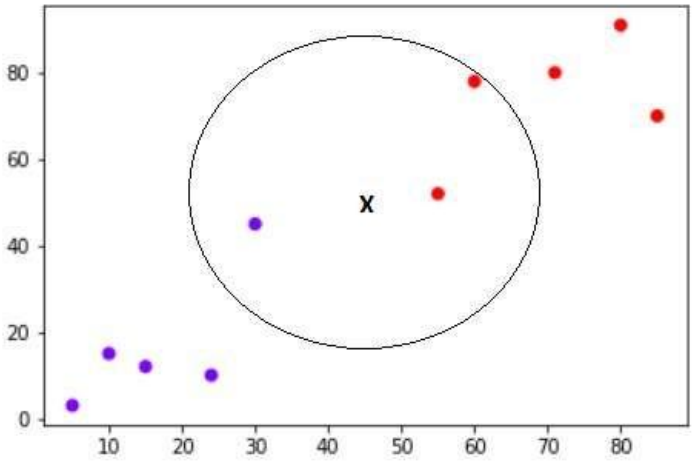

Figure 4. Illustration of KNN

In this algorithm we are predicting whether a transaction is fraud based on the balance update done in the real time. We need to restrict the transactions to K number of transactions to end up detecting fraud of k classification based on similar transactions. Th method helps in predicting a transaction whether it is fraud or not.

3.5 Extreme Gradient Boost

Extreme Gradient Boost (XGBoost) is advance appropriate gradient boosting library. The gradient boosting (GBM) framework is used and still performs better than GBM.

XGBoost transforms weak learners into strong learners belonging to a family of boosting algorithms. Trees grown after each other using the information from the last added tree this shows boosting is a sequential process where weak learners carry out a bit better than random guessing. The data is learnt slowly and helps to better the prediction successive emphasis. Below is a typical example of classification. (Jianrong et al., 2018)

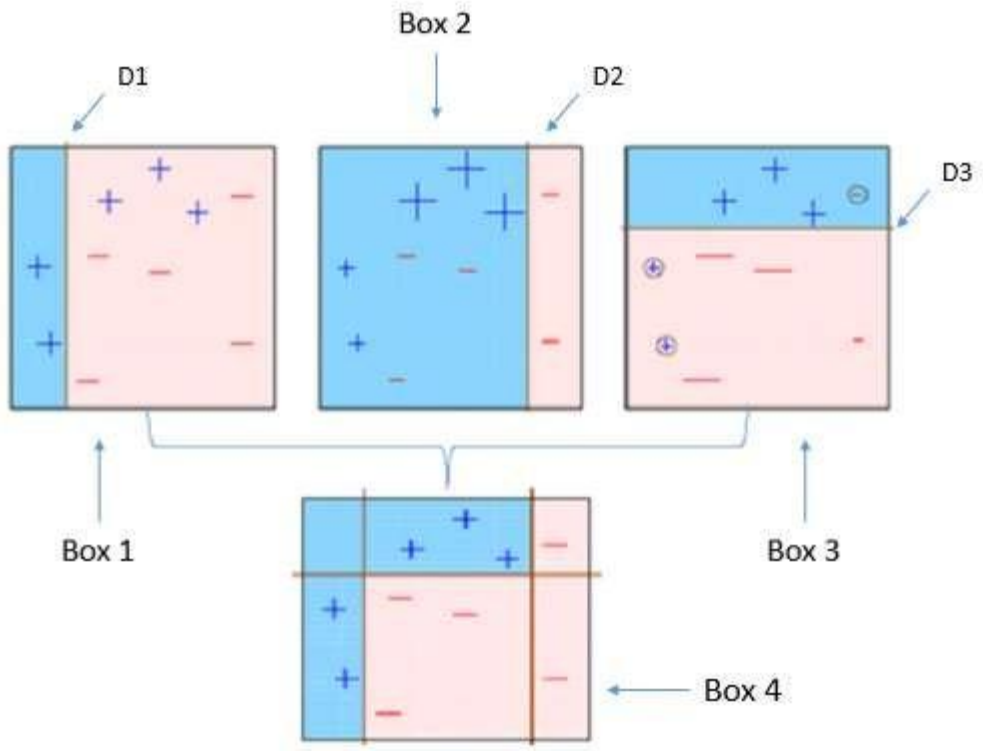

Figure 5. Illustration of XGBoost.

There are four classifiers in four boxes in the above figure and it is difficult to classify + and – most likely homogeneous.

1. Box 1: The first box builds a vertical line (splitting) at D1. The points to the left of D1 is + and rest all to the right should be – but we can see there are three + points misclassified by the classifier.
2. Box 2: The next classifier in Box 2 tries to correct the mistakes. It creates a line D2 and add more weight to misclassified three + points and ti does the same with all the signs to the right of D2 is – and left are + and this causes new errors as there are three- classifiers on the left.
3. Box 3: The Box three again does the same process it contributes more weight to the three incorrect points in the correctly placed points.
4. There is always a classification error associated to each other classifier.
5. These boxes 1,2 and 3 are weak learners and now these classifiers will help in implementing strong classifier Box 4.
6. Box 4: This classifies points correctly as it has combination of the weights of weal classifiers and does good job to correctly classify the points.

This is how boosting algorithms work. The model take advantage and tries to remove the error/misclassification of the last model.

XGBoost can be used for both classification and regression models. Separate methods have been enabled to solve various problems.

3.5.1 XGB Approach

XGb is proposed as a classification model in our research to detect a transaction and predict whether the transaction is a normal transaction or a fraud. The selection of this algorithm is done by selecting important features that help enhance the prediction of algorithm. This model convert slow earner to fast learner in our research. It is proposed to be best suitable model for the fraud detection. The gradient boosting framework helps in improving the model performance. As there are no parameters required in this model it can have fast prediction and fast accurate result. The XGB is evaluated by fitting the target variable which does the prediction and shows if a transaction is fraud or non-fraud.

4. Implementation

The effective analysis of the data would be based on effective implementation of the fraud detection system. The CSV format file is used and Python Language is used for coding with Jupyter Notebook for effective analysis. The simplification of data and refining of the collected data will grant effective analysis.

The following steps are analysed in the continuous process:

4.1 Exploratory Data Analysis:

In exploratory data analysis (EDA) we describe the dataset in brief and implement various visualizations of the data to identify the main characteristics to be focused on.

4.1.1 Synthetic Financial Dataset for Fraud Detection

This dataset was a sample of a much larger dataset generated from a simulation that closely resembles the normal day-to-day transactions including the occurrence of fraudulent transactions. The data is collected from Kaggle: https://www.kaggle.com/ntnu-testimon/paysim1

The dataset was made for performing research on fraud detection methods.

This is a sample of 1 row with headers explanation:
1, PAYMENT,1060.31, C429214117,1089.0,28.69, M1591654462,0.0,0.0,0,0 The dataset explanation:

- step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
- type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
- amount - amount of the transaction in local currency.
- nameOrig - customer who started the transaction
- oldbalanceOrg - initial balance before the transaction
- newbalanceOrig - new balance after the transaction
- nameDest - customer who is the recipient of the transaction
- oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
- newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
- isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.

- isFlaggedFraud - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

4.1.2 Data head

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

Figure 6. Dataset

1. The data has no missing values.
2. The data consist of more than 6 million observations.
3. It has 11 variables.
4. Maximum transactions have amounts less than 1 million euros.
5. Most observations in the dataset are of valid transactions, so any patterns related to identifying fraud transactions may be hard to see, data is also unbalanced.
6. From the sample of observations, there are many instances where what happens to the recipient account (oldbalanceDest, newbalanceDest) does not make sense (e.g. the very first observation involved a payment of 9839.64 yet, the balance before and after the transaction equals 0.)

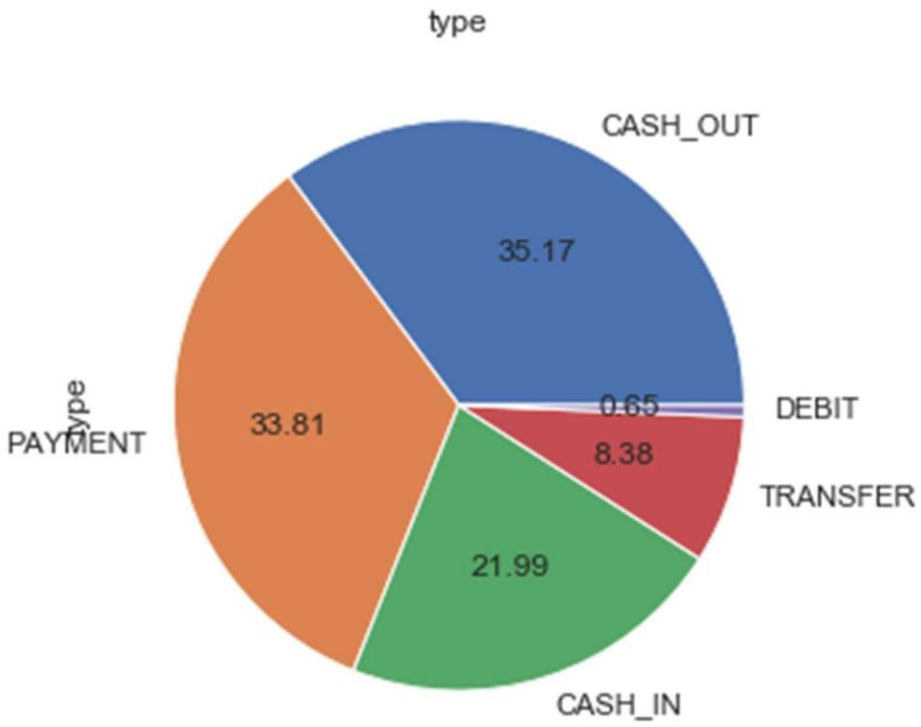

Figure 8. Pie Chart of types of transaction.

The above pie chart is shows types of transaction in the financial dataset. It is data visualization implemented in python matplotlib library. The pie chart is divided into five types of transactions namely CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER. The figure shows that maximum transactions happen as CASH_OUT which is 35.17% and the least transactions happen as DEBIT which is 0.65. We will further explore the dataset in the next visualization.

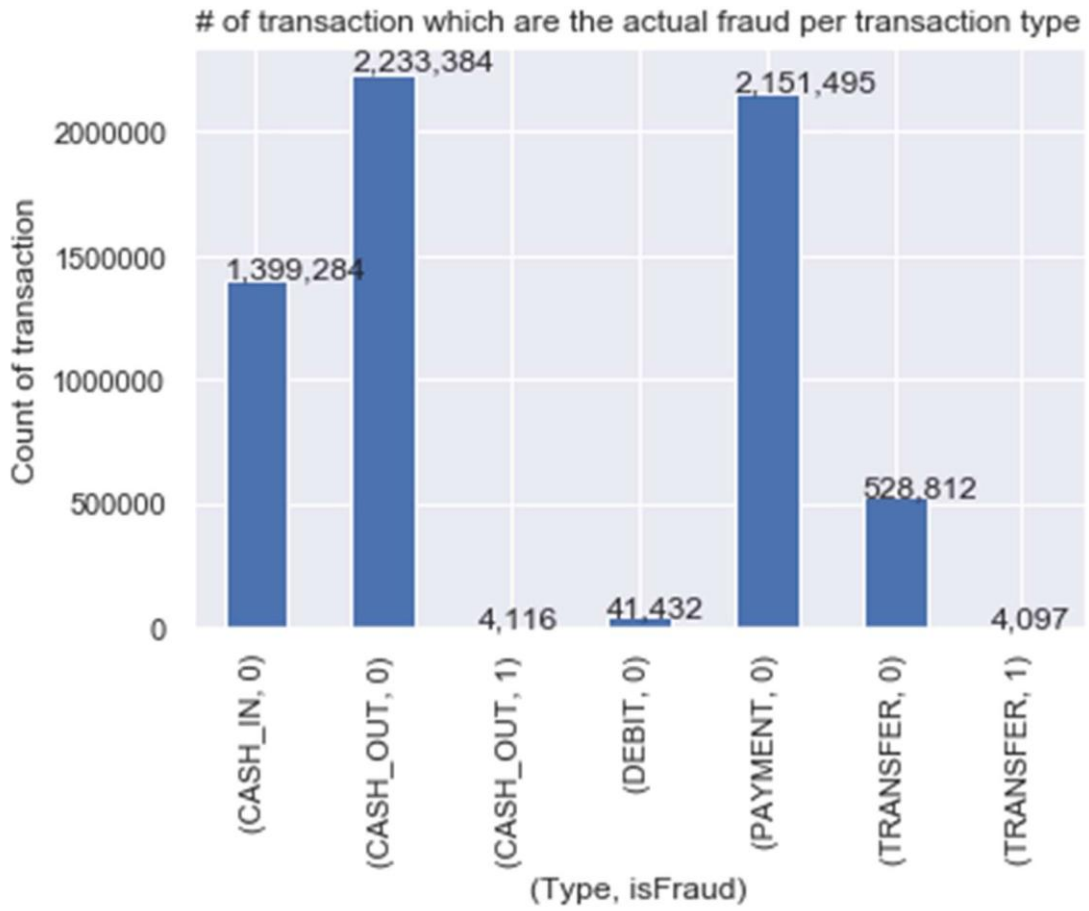

Figure 9. Bar plot of Count of Transactions vs isFraud.

This bar plot shows the count of transactions based on whether a transaction is fraudulent or non-fraud. The plot is build in python matplotlib library. The plot describes the count of transactions where CASH_OUT and PAYMENT are the most used types of transactions and we can see that the DEBIT is the least occurring transaction. If a transaction is fraud it is indicated as 1 and when it is not fraud it is indicated as 0. The fraud detection count is displayed with the transaction type where fraud happens on in the amount going away from the account. The types of transaction which are prone to fraud are CASH_OUT and TRANSFER.

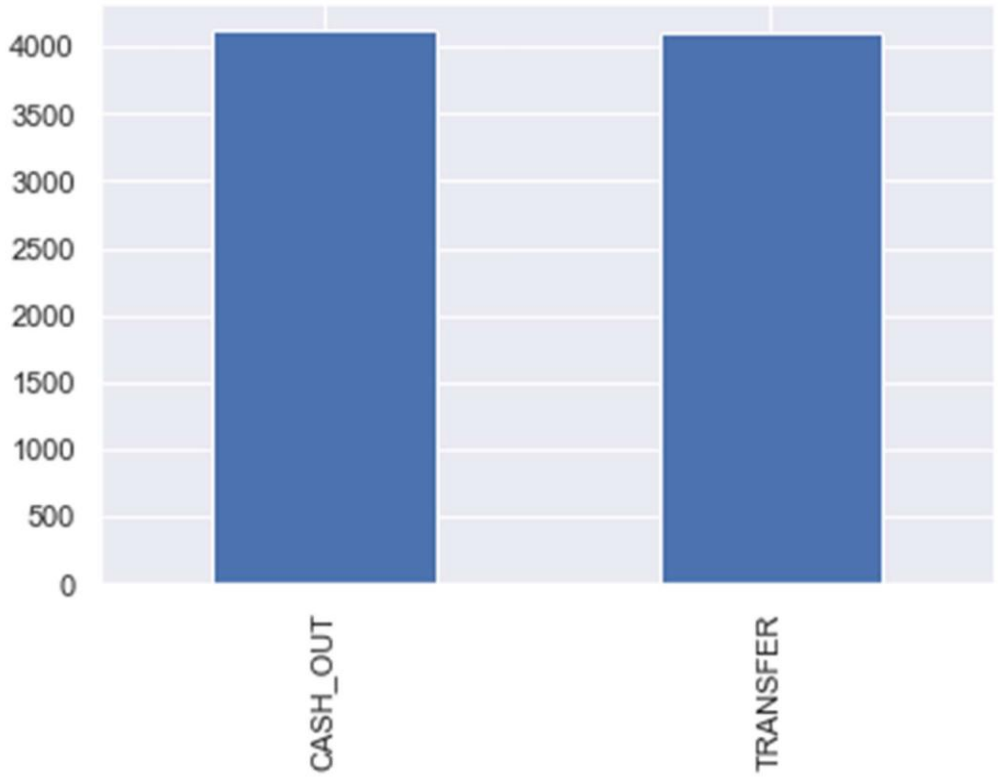

Figure 10. Bar Plot: Count of types of Transaction which are Fraud.

The simple bar plot shows number of fraud transactions based on count of fraud transactions indicated as 1 in the previous visualization. This plot is implemented in matplotlib library of python programming. The plot shows that a total of 8213 transactions are actually fraud out of which 4116 transactions are fraud in CASH_OUT type of activity and 4097 number of transactions are displayed as fraud in the TRANSFER type of transaction. In the next visualization we will se how many transactions are detecting as fraud compared to the actual 8213 fraud transactions.

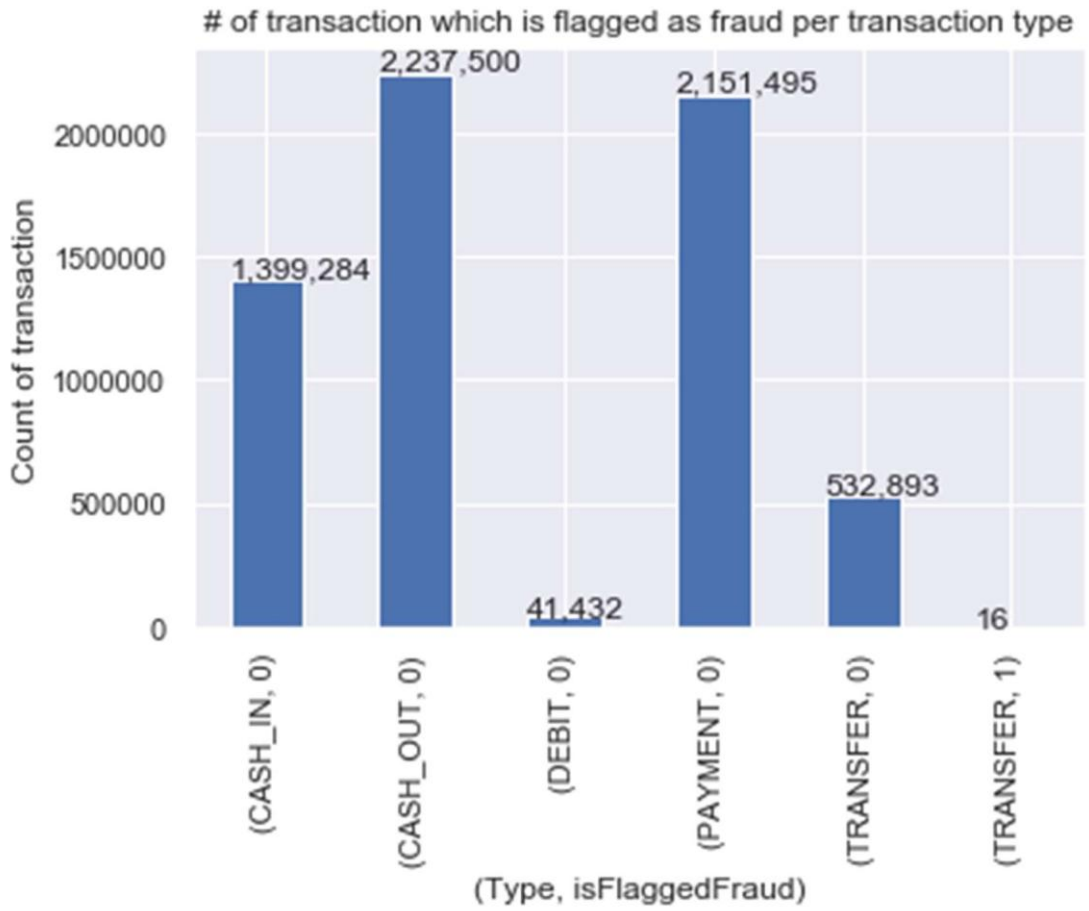

Figure 11. Bar Plot: Count of Transactions vs isFlaggedFraud

The plot is visualized to display the number of transactions which are flagged as fraud based on the type of transactions. The plot is constructed in the matplotlib library of the python language. As we know from the previous visualization that there are 8213 actual fraud transactions. We can see in this graph where only 16 transactions are flagged as fraud in the transaction type TRANSFER indicated as 1. So the system cannot detect 7997 transactions.

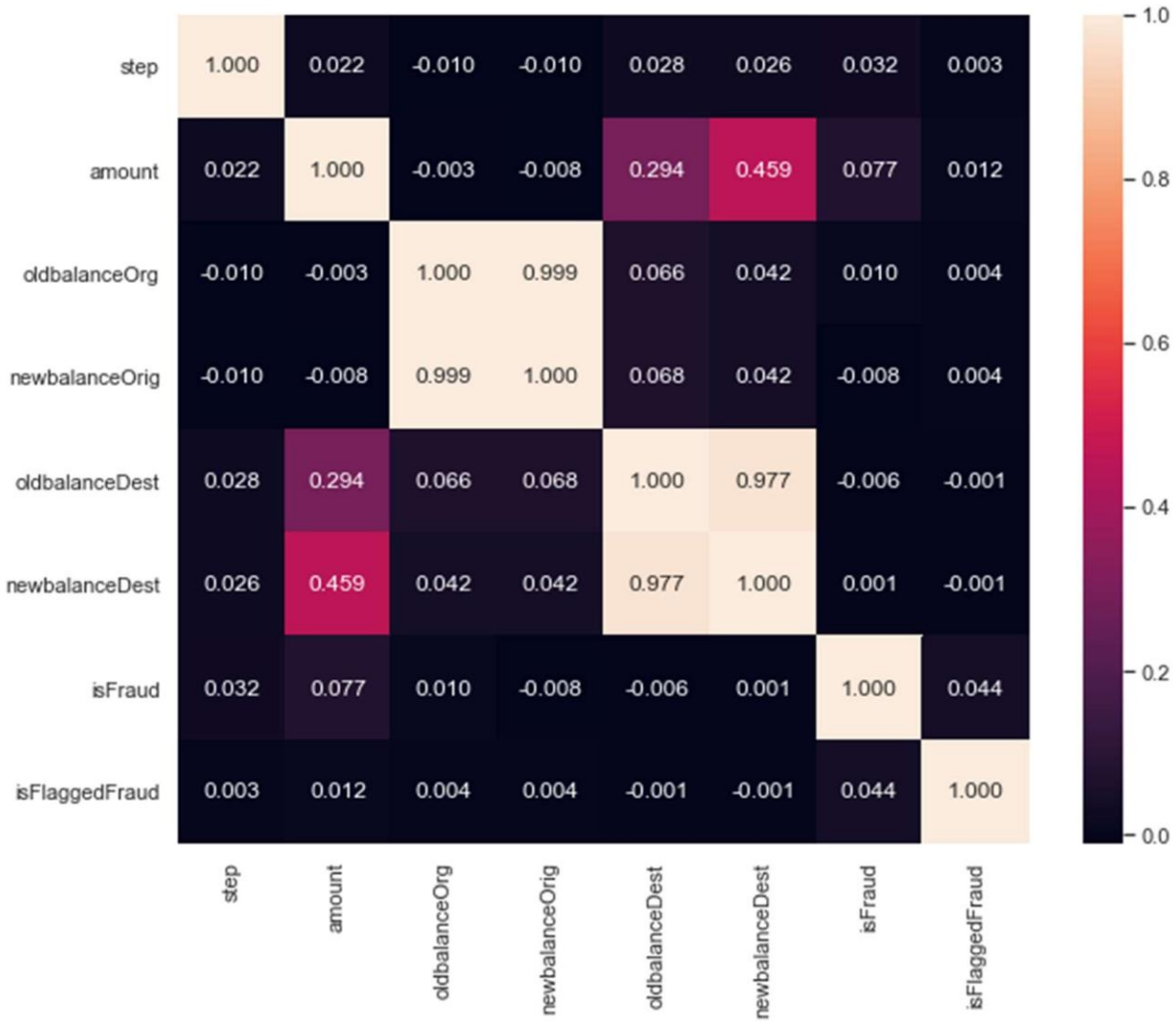

Figure 12. Correlation Heatmap of Dataset Features.

The heatmap shows correlation of the dataset based on the feature content. The visualization is created in the seaborn package of python language. In this heatmap we can see that the features are highly correlated. The the highest correlation is between the couple of features newbalanceDest and oldbalanceDest , newbalanceOrig and oldbalanceOrig. Hence after observing such high correlation we can drop either one of two features for each one. So we remove the oldbalanceOrig and oldbalanceDest.

Figure 13. Confusion matrix of Amount vs isFraud.

After balancing correlation between the features of the dataset we will try to balance the features we need to implement. The matrix is visualized using the seaborn package of python language. The above correlation matrix shows correlation between the transaction amount and fraud statement. Based on this matrix we modify the columns for better prediction of the target variable.

4.2 Data preparation:

Data preparation is one of the important steps in order to have correct methodology. Alteration of data for improving the performance by performing cleansing and removing not required data such as NA, missing labels, missing values etc also balance the data as per requirements.

In the dataset there are many categorical variables which need to be converted as most of the variables used in computation of algorithms are numbers. For this we need to encode the categorical variables with string int numbers. The type of variables are with no hierarchy or order, thus the numerical variables which we use in the categorical variables is call 1 hot encoding. One hot encoding is category of categorical variable which creates indicator variables.
The indicator category variable works with 0 and 1. So if a particular transaction of a particular category e.g. transaction type is TRANSFER) the indication with the associated

with that category will be 1 and if it is not part of the specific category then indicator variable would be 0.

	step	amount	newbalanceOrig	newbalanceDest	isFraud	isFlaggedFraud	CASH_OUT	DEBIT	PAYMENT	TRANSFER
0	1	9839.64	160296.36	0.0	0	0	0	0	1	0
1	1	1864.28	19384.72	0.0	0	0	0	0	1	0
2	1	181.00	0.00	0.0	1	0	0	0	0	1
3	1	181.00	0.00	0.0	1	0	1	0	0	0
4	1	11668.14	29885.86	0.0	0	0	0	0	1	0

Figure 7. Handling Categorical Variables

As we can see in the above snippet we drop tables nameDest and nameOrig as they are not useful in the feature selection.

4.3 Feature Engineering:

4.3.1 Random Under Sampling:

Under sampling of the imbalanced data is very important as we can see in the plot below the target variable isFraud is highly imbalanced with 6,354,407 number of transactions are not fraud which is indicated as 0 and 8213 are marked as fraud and indicated as 1. So for balancing of data we need to apply sampling technique to our dataset. We try to balance the data set by applying the random under sampling technique. This technique tries to balance the data by reducing the class with majority samples i.e. isFraud with indication 0 randomly. This process takes place until data is not balanced to the 50:50 ratios of fraud and non-fraud transactions. When the technique is implemented we have a sub set of our data with a 50:50 ratio of the regarding classes. In the next step the balanced data is used for selection of models. (Dahee & Kyungho, 2017)

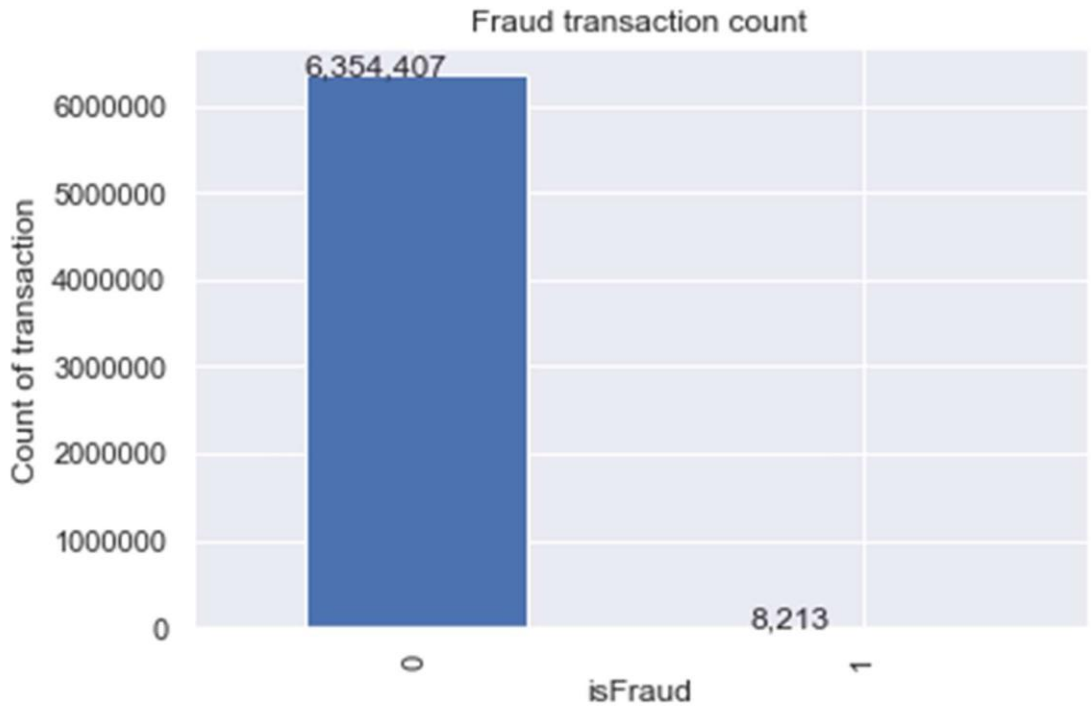

Figure 22. Count of Transactions vs isFraud[0,1].

4.4 Evaluation of Model:

With all the above points from dataset to data preparation and after balancing of the imbalanced dataset we cautiously examine various algorithms with their accuracy and predictions. We decide to apply 5 machine learning models which are Logistic Regression, Random Forest, Decision Tree, XGBoost and K Nearest Neighbors.

4.4.1 Selection of Model: XGBoost

We have implemented the machine learning algorithms by using python programming. The libraries which were required to be implemented are imported and we have added a figure of how the model has been evaluated. On successful importing the modules we do selection of features for the model prediction from independent variables and dependent variables and we split the dataset with ratio of 70:30 training and testing data. After splitting the data with the specified ratio the selected model is applied XGBoost on various set of splits and the performance of the model is observed. Henceforth we do check the results and performance evaluation of the model used on the dataset. The prediction is made of the column isFraud where the data predicts whether a transaction is fraud or non-fraud.

4.4.2 Selection of Model: KNN

Similar technique is used to evaluate KNN based on the results which we require from this model. Superior models are those whose reality is presented accurately, being preferably to the point and accountable. We have trained the model and applied the required feature engineering. We apply the required parameters for tuning of the model and fitting the model. The model selection is done with the proposed target variable. Then we validate the model and approve the selection.

4.4.3 Selection of Model: Random Forest

The figure attached below is in python and we have imported the required libraries and dataset for the model evaluation which is performed in Jupyter Notebook. When everything is imported successfully and feature selection is done on dependent variable and independent variable and the dataset is divided with ratio 70:30 for train and test. After dividing the dataset into derived ratio Random Forest is applied with different hyperparameters for tuning and various split and validated. The prediction of data is done and the results are displayed as accuracy, confusion matrix and classification report. The prediction is dependent on the column isFraud where we predict if the transaction is non-fraud or fraud.

4.4.4 Selection of Model: Logistic Regression

Logistic Regression is applied for classification problem. It is supervised algorithm. We used linear model to get the prediction of the target variable which shows fraud and non-fraud transactions. The classification is performed on the dataset separating into test and train sets with ratio 70:30 as we find the ratio best fit for the model. We used random under sampling technique and used the sample data for prediction of train and test set. The fraudulent prediction is implemented on the test set and results such as accuracy, confusion matrix and classification report are displayed.

4.4.5 Selection of Model: Decision Tree

The implementation of the model is done in python programming using Jupyter notebook. We have imported all the necessary libraries required for the evaluation of the model. After selecting all the necessary imports we do feature engineering where we apply random under sampling. This balances the target variable features and it helps in improving the performance of the model. The variables are selected and the data is split into desired ratio of 70:30 which

is decided with critical observation. When data is split in train and test set we apply Decision Tree algorithm with the specific parameters for best tuning of the model validation. The model prediction is evaluated and the outcome is recorded as confusion matrix, accuracy and classification reports. The model predicts the detection of fraudulent case whether the transaction is fraudulent or not.

4.5 Performance Measures:

4.5.1 Confusion Matrix:

Classification of a problem based on prediction results is known as confusion matrix. Summary of prediction is based on number of incorrect and correct categories with de-escalation and count values. The prediction shows the ways the model is confused in different was of classification which is the key to confusion matrix. They give us detailed insight of what errors are made by classifier and most essentially which type of errors are being made by the classifier.

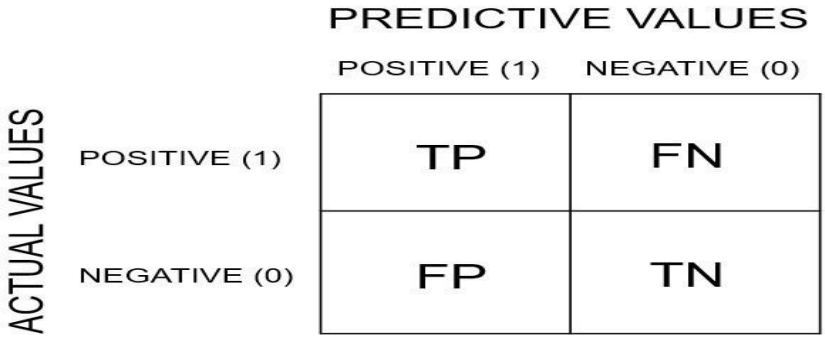

Figure 14. Confusion Matrix Illustration

In the above figure,
Column 1 predicts Positive Values
Column 2 predicts Negative Values
The terms of the figure are defined as:
Positive (1): It shows the class is fraud.
Negative (0): It shows that class is not fraud.
TP (True Positive): This states that class is positive and prediction is positive.
FN (False Negative): This states that class is positive but prediction is negative.
TN (True Negative): This states that class is negative and prediction is negative. FP (False Positive): This states that class is negative but prediction is positive.

4.5.2 Accuracy:

Accuracy is a ratio of correctly predicted observations to the total observations which is the most important performance measure of the model. We can say if the accuracy of a model is high then the performance of the model is best. Accuracy is really good measure but that only applies on symmetric datasets when we can say that the values of false negatives and false positive are almost similar. Hence if the previous sentence is not true then we need to evaluate other performance measures for the model.

4.5.3 Recall:

The recall is based on the full range of truly classified positive examples and total set of positive examples which is quantitative relation. The class is perfectly recognized if it has high recall values with very few False Negative examples and it is given as:

4.5.4 Precision:

Precision is based on the values of whole range of predicted positive examples which are divided by the entire range of properly classified positive examples. The class is recognized as positive if it is tagged as positive with a few range of False Positive and it is given as:

4.5.5 Specificity:

The proportion of negative (or true negative) that are foretold for the proportion of actual negatives is specificity. This means there is another proportion of negative values which turned as false positive which are supposed to be actual negatives. The false positive rate is the term defined by this proportion. Total of false positive rate and specificity would be one.

4.5.6 F1 Score:

F1 Score is based on both false positives and false negatives and weights the average of Precision and Recall. F1 score is not as easy to understand as accuracy but it is more effective than accuracy, particularly if there is uneven distribution.

Accuracy works fine if there is a similar cost for false positive and false negative. If there is huge difference between false negative and false positive then it recommended to look on Recall and Precision both the measures.

4.5.7 AUC:

Area under the curve (AUC) is used for prediction of the models where its shows
which classes work better in the classification analysis. Evert potential classification thresholds in AUC are provided by merging performance and measures. The chance that model ranks random positive example very easily than a random negative positive example because of a decoding technique in the AUC.

4.5.8 ROC curve:

Receiving operating characteristic (ROC) curve has the ability of binary classification system which plots a graphical illustration of the model. It describes difference between the relative performance and suitability of the machine learning model. To a particular dataset it has a feature which gives best model prediction. The figure attached below is a snippet of ROC curve for a specific model and analysis for best model suitable based on the curve.

5.Results and Discussion

This section is all about the results of our proposed model we discuss the prediction of the machine learning models we have applied in our implementations. We have displayed some collective graphs to support our discussion using the matplotlib library of the python programming.

5.1.1 Model Accuracy, Confusion Matrix, Classification Report and ROC Curve: Logistic Regression

```
=== Model Accuracy ===
0.6879058441558441


=== Confusion Matrix ===
[[1162  221]
 [1317 2228]]
```

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.84      0.47      0.60      2479
           1       0.63      0.91      0.74      2449

   micro avg       0.69      0.69      0.69      4928
   macro avg       0.73      0.69      0.67      4928
weighted avg       0.73      0.69      0.67      4928
```

```
=== All AUC Scores ===
AUC: 0.69
```

Logistic Regression is fitted by depicting the above information results. The results of Logistic Regression show that it is the worst performing algorithms from the picked algorithms. It has low performance than all other models. The accuracy is 68.7%. The rest of the measure such as confusion matrix and classification reports with measure precision, recall, f1-score and support all have similar poor performance as the accuracy of the model.

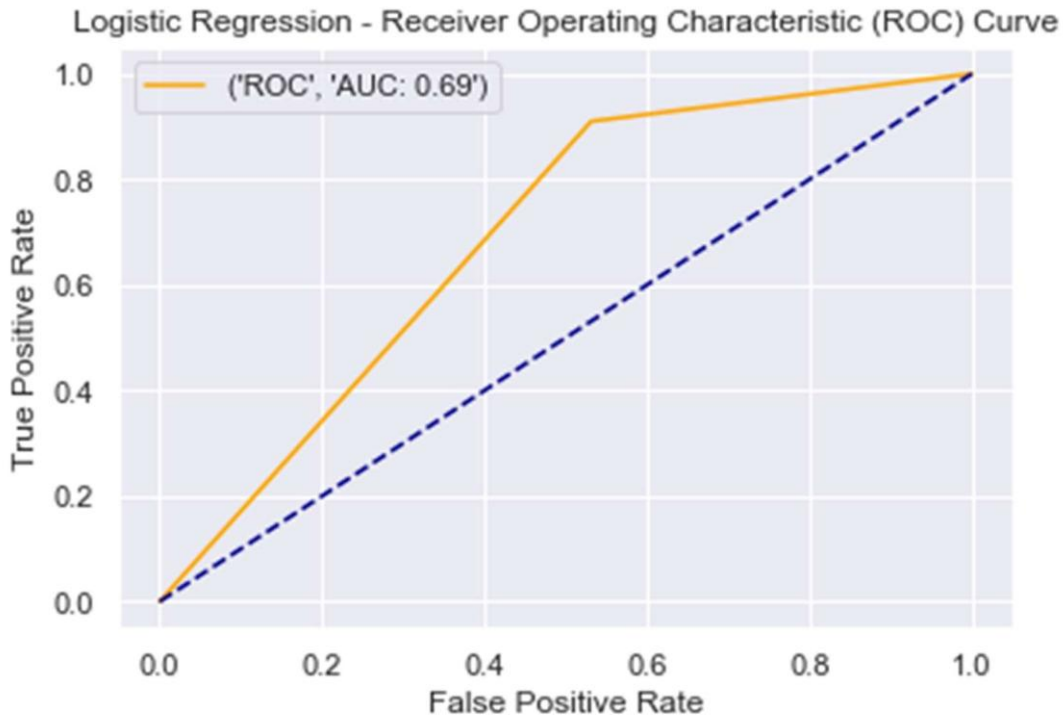

Figure 15. ROC Curve – Logistic Regression.

5.1.2 Model Accuracy, Confusion Matrix, Classification Report and ROC Curve: Random Forest

```
=== Model Accuracy ===
0.929788961038961


=== Confusion Matrix ===
[[2324  191]
 [ 155 2258]]
```

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      2479
           1       0.94      0.92      0.93      2449

   micro avg       0.93      0.93      0.93      4928
   macro avg       0.93      0.93      0.93      4928
weighted avg       0.93      0.93      0.93      4928


=== All AUC Scores ===
AUC: 0.93
```

The result of this model is fitted by applying Random Forest algorithm. The result consist of accuracy which is close to 93% which is overall one of the finest among the algorithms compared. The overall results in better than other proposed algorithms. In the confusion matrix it has the least number of valid transactions predicted as fraudulent transactions. The classification report has the overall result where the performance of the model has been best compared to others based on measures recall, precision, f1-score and support.

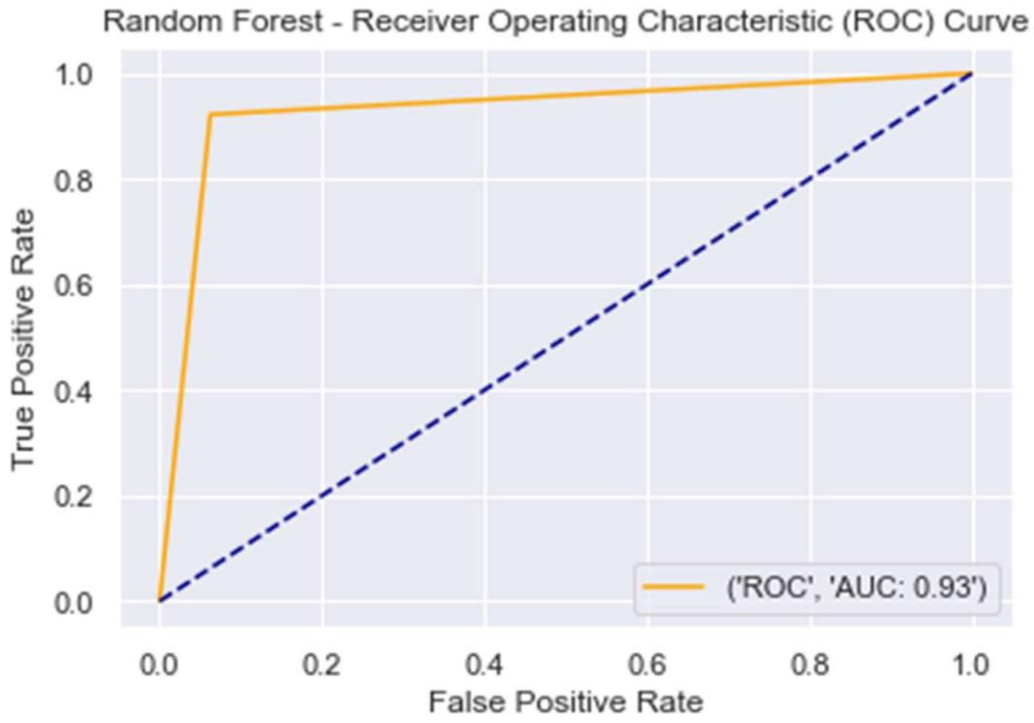

Figure 16. ROC Curve – Random Forest.

5.1.3 Model Accuracy, Confusion Matrix, Classification Report and ROC Curve: Decision Tree

```
=== Model Accuracy ===
0.9281655844155844


=== Confusion Matrix ===
[[2372  247]
 [ 107 2202]]
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.91      0.96      0.93      2479
           1       0.95      0.90      0.93      2449

   micro avg       0.93      0.93      0.93      4928
   macro avg       0.93      0.93      0.93      4928
weighted avg       0.93      0.93      0.93      4928



=== All AUC Scores ===
AUC: 0.93
```

This model is fitted by applying Decision Tree algorithm. The result consist of accuracy which is close to 93% which is overall one of the finest among top three the algorithms compared. The overall results in better than other proposed algorithms. In the confusion matrix it has the one of the least number of fraudulent transactions predicted as valid transactions. The classification report has the overall result where the performance of the model has been best compared to others based on measures recall, precision, f1-score and support.

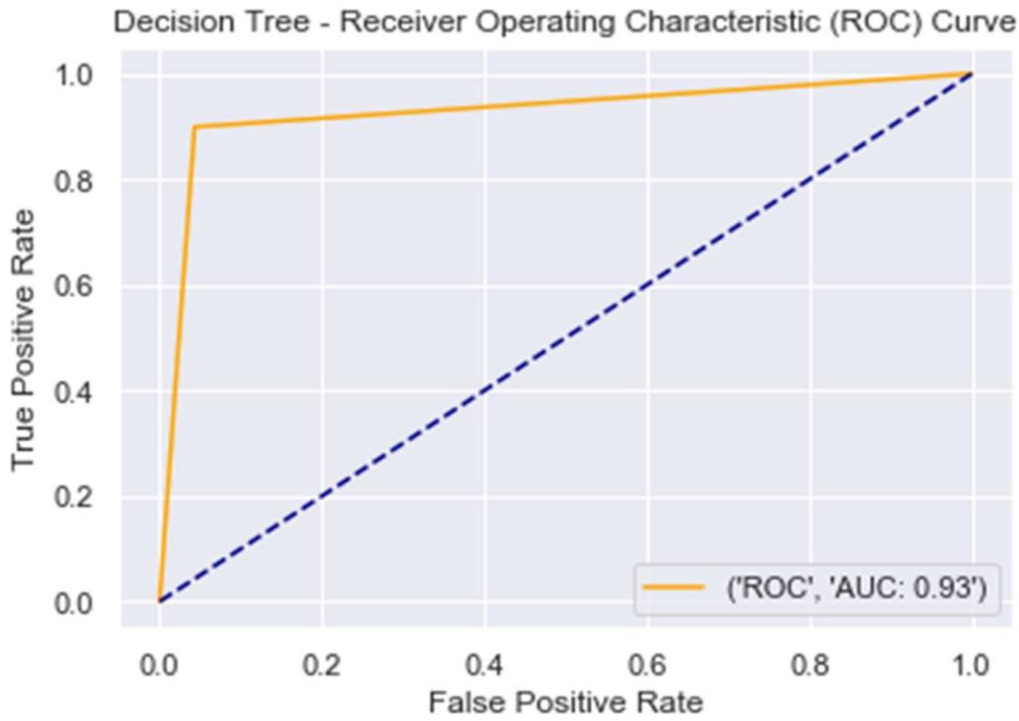

Figure 17. ROC Curve – Decision Tree.

5.1.4 Model Accuracy, Confusion Matrix, Classification Report and ROC Curve: Extreme Gradient Boost

```
=== Model Accuracy ===
0.9318181818181818


=== Confusion Matrix ===
[[2374  231]
 [ 105 2218]]
```

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.91      0.96      0.93      2479
           1       0.95      0.91      0.93      2449

   micro avg       0.93      0.93      0.93      4928
   macro avg       0.93      0.93      0.93      4928
weighted avg       0.93      0.93      0.93      4928


=== All AUC Scores ===
AUC: 0.93
```

XGBoost is the model with the best accuracy compared to all the proposed techniques. We fit the XGB and displayed the accuracy of 93.18%. The model also has best results in confusion matrix for least number of fraudulent transactions predicted as valid transactions. Numbers say that XGB is the superior model among the ones selected in the research. The classification report has the overall result where the performance of the model has been equally good compared to others three top models based on measures recall, precision, f1-score and support.

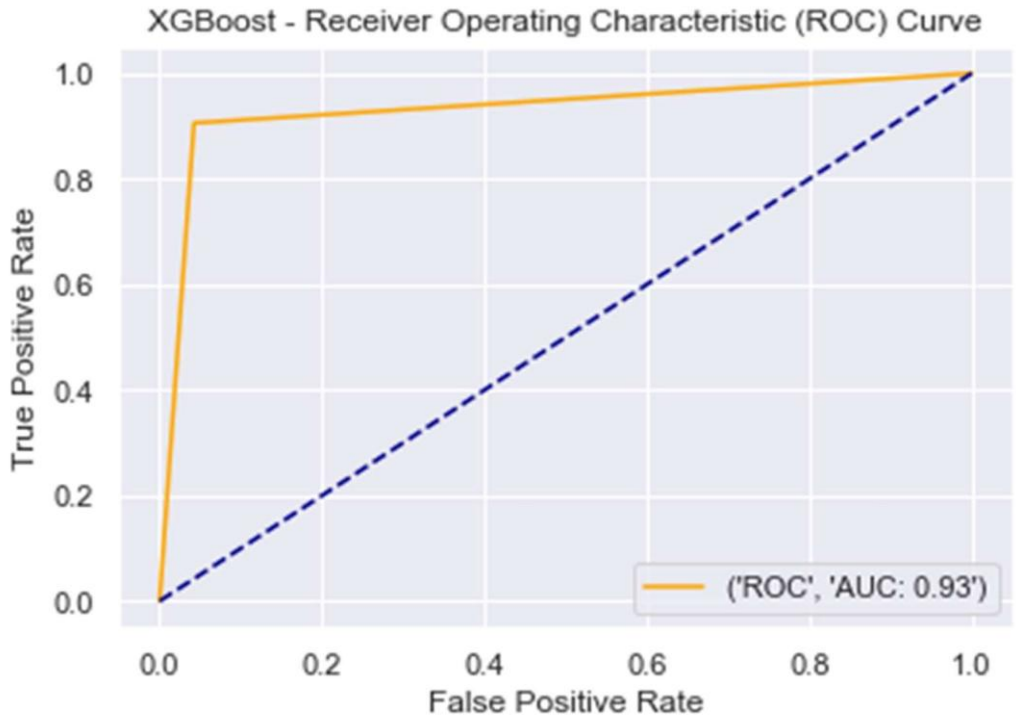

Figure 18. ROC Curve – XGBoost.

5.1.5 Model Accuracy, Confusion Matrix, Classification Report and ROC Curve:

K Nearest Neighbor

```
=== Model Accuracy ===
0.8238636363636364


=== Confusion Matrix ===
[[2204  593]
 [ 275 1856]]
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.79      0.89      0.84      2479
           1       0.87      0.76      0.81      2449

   micro avg       0.82      0.82      0.82      4928
   macro avg       0.83      0.82      0.82      4928
weighted avg       0.83      0.82      0.82      4928



=== All AUC Scores ===
AUC: 0.82
```

The KNN model fitting is done by using the target variable which predicts fraud or non fraud. We inspect the accuracy of the KNN algorithm which come to 82% and is considered as underperforming model as we have three other models with great accuracy. The result is also supported by confusion matrix, classification report with measures such as f1-score, precision, recall and support.

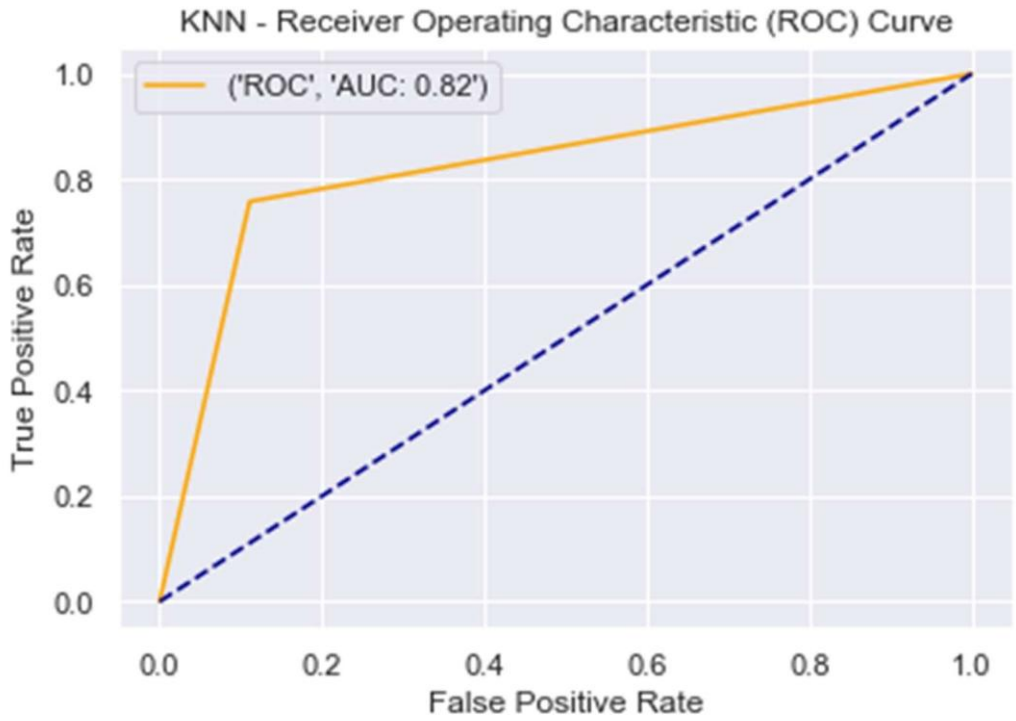

Figure 19. ROC Curve – KNN.

5.2 Algorithm Comparison:

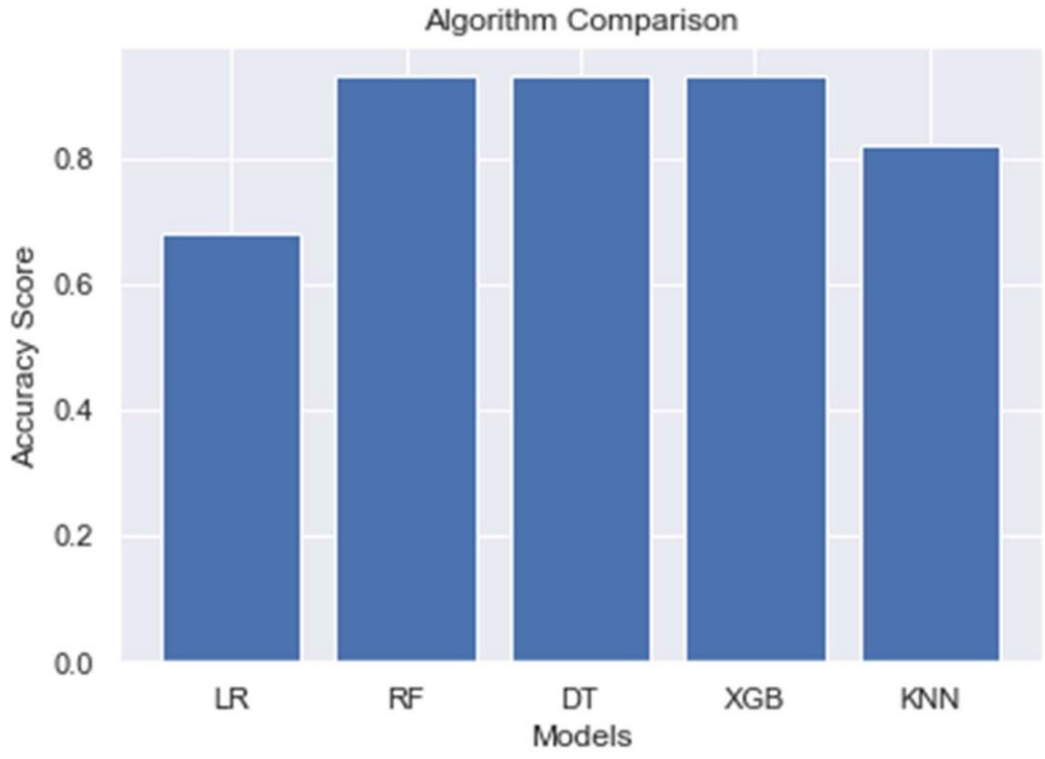

Figure 20. Algorithms Comparison.

The above result plot shows accuracy comparison of various algorithms we have used in the implementation of our research. The bar plot is build in matplotlib library of python language. As we can see that five classification algorithms are applied in the proposed work and we have the results of all five models. In the graph Random Forest, Decision Tree and XGBoost models have same accuracy which is 93% and the other two under performing models KNN with 82% and worst among the considered algorithms Logistic Regression with accuracy of 68% we can say that XGBoost is the best performing model with highest accuracy of 93.2% to but there is negligible difference in the results between the three high performing models.

5.3 ROC Comparison:

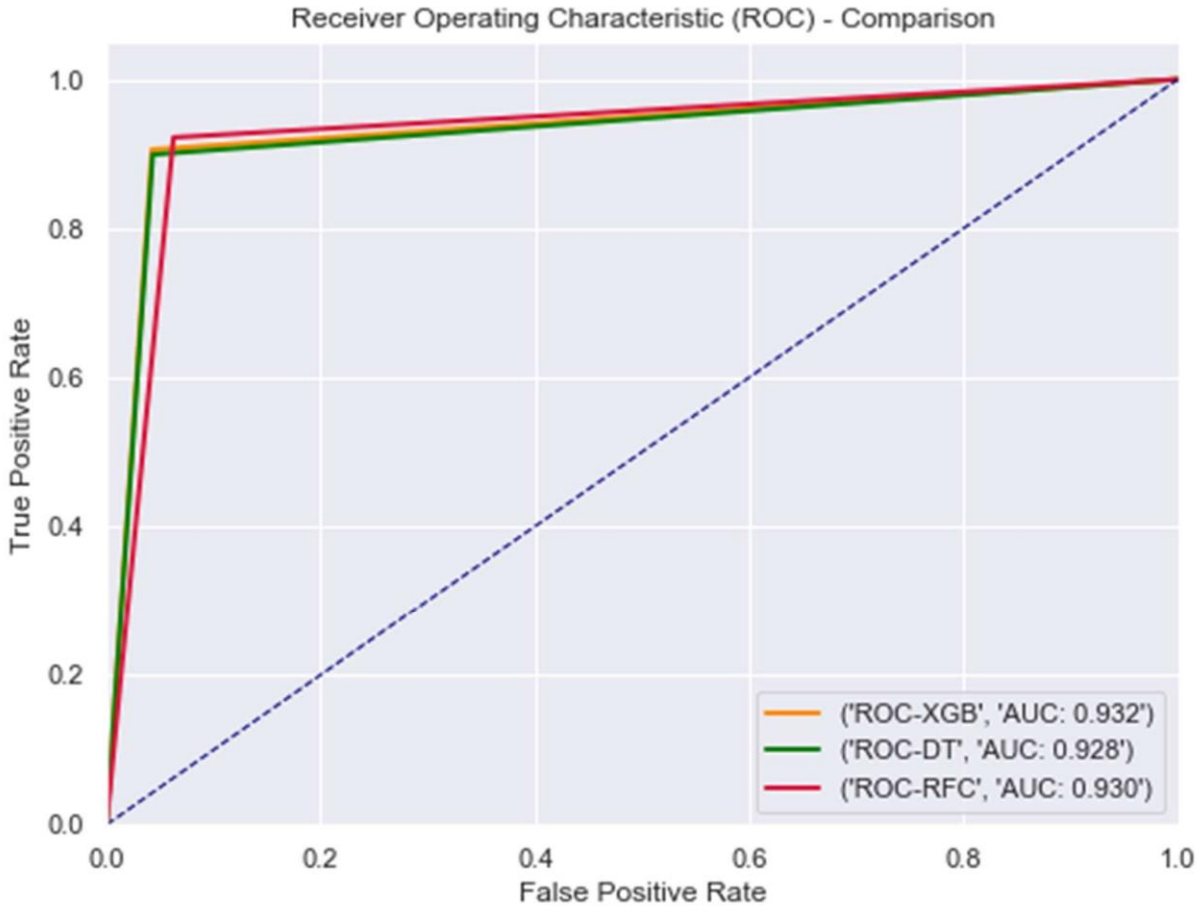

Figure 21. ROC Comparison.

We have selected three best performing models Random Forest, Decision Tree and Extreme Gradient Boosted Trees and displayed their ROC curve to discuss the superior model. While there is no difference in the AUC of both the curves. Both the models are same for every practical purpose both are eventually same. Overall Random Forest would be deemed to be superior model as cost of resolving valid transactions labelled as fraudulent would exceed the cost of resolving a handful more fraudulent transactions that have been passed off as valid.

6.Conclusion and Future Work:

In this project we present a financial fraud detection problem on a mobile money transaction dataset. The machine learning techniques used are supervised learning. We have applied binary classification on our model. We have carefully decided to apply five popular classification machine learning model including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K Nearest Neighbor (KNN) and XGBoost (XGB) to solve this problem. The dataset being a simulation data and a public data PaySim it displays that three out of the five proposed models perform satisfactorily well. XGBoost model should be the first choice based on the practical results, however Random Forest is the most efficient method as it is less cost effective and better performing where the accuracy results of RF, DT and XGB are almost the same. As XGB required no hyper-parameters to tune in our method, where RF and DT need some parameters to be tuned for achieving the optimal results. The other two models KNN and LR perform relatively lower than these three top models in which KNN was tuned with parameters and LR does not have any parameters. As we evaluate the three better performing models RF, DT and XGB they analyze the data with great accuracy and give deep insights and results of the valid transactions and the fraud transactions with reduced dimension capacity. But the data challenge still remains same in the topic of our research. By applying different techniques only on one dataset is not a sound comparison.

For the future work we hope to work on more dataset for trial. The five methods we described in this research are quite typical, nonetheless they performed relatively good. On the new type of data these models will not be effective, we will need to develop more advanced methods. Fraud Detection is a vast topic there are many types of fraud and many ways to detect it. We would leave this topic for the future work. This problem attracts more of commercial industries than the academic institutions. The main reasons of the deficiency of public data for scientific and systematic comparison is that commercial companies having the data with confidential issues. If more data is available for research, it would attract more in depth research in the field with abundance of interest.

7.References:

Adeyinka, A., Stelios, K., Miltos, P., & Emmanouil, P. (2016). EvaluatingCase-Based Reasoning Knowledge Discovery in Fraud Detection. Sustainability, 11(1570), 31.

Bian, Y., Cheng, M., Yang, C., Yuan, Y., Li, Q., Zhao, J. L., & Liang, L. (2016). FINANCIAL FRAUD DETECTION:
A NEW ENSEMBLE LEARNING APPROACH FOR IMBALANCED DATA. Association for Information Systems AIS Electronic Library (AISeL).

CLIFTON, P., VINCENT, L., KATE, S., & ROSS, G. A Comprehensive Survey of Data Mining-based Fraud Detection Research.

Dahee, C., & Kyungho, L. (2017). Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System. IT CoNvergence PRActice (INPRA), 5(4).

Dongsong, Z., & Lina, Z. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS,
34(4).

G., A., Arun, S. D. P., G.S., R., B.Lalitha, B., K., E., & D., R. (2009). Financial Statement Fraud Detection by Data Mining . Int. J. of Advanced Networking and Applications, 1(3), 159-163.

Jarrod, W., Maumita, B., & Rafiqul, I. Intelligent Financial Fraud Detection Practices: An Investigation.

Jianrong, Y., Jie, Z., & Lu, W. (2018). A Financial Statement Fraud Detection Model Based on Hybrid Data Mining Methods. Internantional Confernece on Artificial Intelligence and Big Data.

Jianrong, Y., Yanqin, P., Shuiqing, Y., Yuangao, C., & Yixiao, L. (2019). Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach. Sustainability, 11(1579), 17. doi:doi:10.3390/su11061579

Kunlin, Y. (2018). A Memory-Enhanced Framework for Financial Fraud Detection. IEEE International Conference on Machine Learning and Applications.

Mahdi, O., Qingfei, M., Vahab, M., & Muhammad, P. (2019). The Efficacy of Predictive Methods in Financial Statement Fraud. Hindawi, Discrete Dynamics in Nature and Society, 12. doi:https://doi.org/10.1155/2019/4989140

Prabin, K. P. (2011). A Framework for Discovering Internal Financial Fraud using Analytics. International Conference on Communication Systems and Network Technologies.

Prasad, S., Shuhao, Z., & Yibing, Q. (IEEE Systems and Information Engineering Design Symposium). Detection of Fraudulent Financial Reports with Machine Learning Techniques. 2015.

QIAN, L., TONG, L., & WEI, X. (2009). A SUBJECTIVE AND OBJECTIVE INTEGRATED METHOD FOR FRAUD DETECTION IN FINANCIAL SYSTEMS. Proceedings of the Eighth International Conference on Machine Learning and Cybernetics.

Rasa, K., & Zivile, G. (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. Procedia - Social and Behavioral Sciences, 213.

Ratha, & Pech. (2019). Fraud detection in mobile money transfer as binary classification problem. .

Sadagali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. ScienceDirect, 148, 45-54.