DESIGN AND EVALUATION OF AN EFFICIENT CLASSIFICATION TECHNIQUE FOR CREDIT CARD FRAUD DETECTION

R Raja Subramanian
Dept of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Krishnankoil, Virudhunagar, Tamil Nadu, India
rajasubramanian.r@klu.ac.in

M.Venkatesh
Department of Computer Scienceand Engineering,
Kalasalingam Academy of Research and Education,
Virudhunagar, India
9919004186@klu.ac.in

N.Liyaz
Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education,
Virudhunagar, India
9919004188@klu.ac.in

M.Sukumar
Department of Computer Scienceand Engineering,
Kalasalingam Academy of Research and Education,
Virudhunagar, India
9919004166@klu.ac.in

P.Dadavalli
Department of Computer Scienceand Engineering,
Kalasalingam Academy of Research and Education,
Virudhunagar, India
9919004204@klu.ac.in

Abstract*—Credit card frauds are a source of very large losses to the users and in banks. The fraud transactions are increased due to online transactions, the fraud detection has became a challenging and an important issue for banks . The research uses old labelled data to build the suitable model that bifurcate the input transactions as fraudlent or non fraudulent. A comparative and determinative study of three classifiers using supervised ML algorithms – SVM ,logistic regression and random forest is used on datasets to build model with different under sampling ratios to choose the best model for our dataset using machine learning algorithms.

Keywords*—online transactions,Fraud, SVM, random forest, logistic ,regression, under sampling,credit ,machine learning.

I .INTRODUCTION

Due to credit card Fraud many millions of amount is lossed by users every minute as reported in various articles. , Detection and prevention. System studying techniques have Been in use on account that Nineties for fraud detection with the algorithms Becoming increasingly high rate and becoming tough task to these Fraud detection techniques are specifically categorized Into – misuse or fraud detection and anomaly detection . fraud Detection uses old statistics to label a transaction as fraud or Valid. We can either supervised or unsupervised strategies May be used to reap this. Anomaly detection makes use of transaction Facts of customers behaviour to come across any anomaly. The thieves use many ways to do frauds in credit cards these can be prevented only if we find that these are an abnormal taransaction ,the fraudlent transactions of credit card are of many types those can be in internet or by stealing the credit card of users Some criminals use lost or stolen credit cards to commit fraud. Others make illegal transactions without ever having the credit card in their possession. Card-not-present fraud only requires the criminal to know basic card or account details to access the victim's funds .,.Fraud detection is a set of unwanted activities that are taken to prevent money or property from being obtained through false pretenses.These are very tough task to predict for normal human here we use machine learning for these very high computational tasks,for processing the whole data and we make those transactions to classify with the classification algorithms, Frequently checking your credit report to see if anything seems un common,such as new credit searches and inquiries, the opening of new accounts, or your hard-earned credit.

.

Once stolen by cyber criminals , they could use your "card" to Note only make large purchases, but to also empty out your of Account in ajiffy. They couldpurchase something online or at any retail store.

II .*RELATEDWORK*

We can use multiple ways for solving credit cards frauds by using supervised and unsupervised algorithms are used for fraud detection our target is to succeed in dealing with frauds included dataset i.e., strong class imbalance, the included with group of labelled and unlabelled data samples,andtoovercome by the increasing abilitytoprocess alargenumberoftransactions of Money.Different Machine learning algorithms of supervised are classification tree, Naive Bayes Classifer, Logistic Regression ,and Least Squares method,Support Vector Machine, are used to find fraud transactions in the given dataset.In Random Forest Algorithm there are two ways to train the behavioural transaction features of a type normal and abnormal type transactions. so these can be trained by these algorithms, They are Random-forest based random tree .our aim is to focus and overcome the above-mentioned issue by improving the random forest by the technique present in random forest like under sampling the given data. Research on credit card fraud using machine learning techniques for fraud detection has started during in the year ninties with ANN(Artificial Neural Network) being one of the most popular algorithm to be use. And another author utilized the method in neural network trained transaction data *labelled fraud and non-fraud. A three layerfeed-forward neural network was used to evaluate transactions and differentiate them based on a threshold value. Though the model executed best calculation in terms of performance and accuracy, it was limited with high computation time. Author Srivastava.[5] used a HMM (Hidden Makov)for anomaly detection. A HMM rule was initially train with the cardholder's common behaviour.If the HMM model does not accepts the current transaction with sufficiently high Research probability, the transaction is flagged as fraudulent.* However, only transaction amount is considered as a feature in this model .one more author used ID3 and Svm for fraud detection. The data is splitted into three multiple groups with different ratios of models are improved based on these datasets and the algorithm performance was calculated. Results show that ID3 performs better than svm.

As the imbalance of the dataset increases, accuracy metric is not a good performance metric and the model is restricted by its use of accuracy as it overall performance metric. later, random forest algorithm is used to solve the present issue in problem. Here, CART trees based on "bootstrapped" samples of data and then combines the predictionsbased random forest were used to detect fraud. A starting collation of the two was done and CART basis random trees of random forest was picked up due to its better performance. The approach of under sampling is utilised to handle the issue in problem of imbalanced data and choose the good ratio of under sampling.A smallset of the training data set is sampled in a random way so that we train everyindependent branch of treeon which tree is splits in to branches then it becomes decision tree, each and every Decisionnode and chance nodes of tree then splits on a feature designated from a random small sampleset of the complete A large dataset is then utilized to reveal the effectiveness of the algorithm. Though the algorithm depicts best outcome on small amount of data, the effectiveness on highly imbalanced dataset stillremains.

Random forest for feature Selection ,with the huge data sets having more features,training the dataset isextremely veryfastintherandomforest algorithm and because in random forest training is done independent to every tree in random forest ,The Random Forest algorithm is a good way to implement the Stochastic discrimination. This is good to find resistant and error that is generalization error .These makes us to choose random forest as best suited for this credit card fraud detection.

Advantages : Random Forest Algorithm chooses or finds the best suited feature instead of significant feature this type of behavour of random forest tells us as the best Algorithm for the model.

It classifies the output as fraud and non fraud for positive result is fraud and for negative result is non fraud that is binary classification as 0 and 1.

A. Proposed Scheme*:*

In this Proposed Scheme we are finally using the random forest algorithm for the binary classification of the given dataset .this Algorithm suits best for regression and classification , The best advantage of Random forest is that it focus on collecting various decision trees to come at any solution. This is an ensemble algorithm that considers the results of more than one algorithms of the same or different kind of classification..The sklearn.ensemble includes 2 algorithms on bsasis of randomized decision trees: the RandomForest algorithm and the Extra-Trees method. Both algorithms are perturb-and-combine techniques specifically designed for trees. This means a diverse set of classifiers is created by making substitution of randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers..// This indicates that Random Forest algorithm has been good algorithm for overfitting as well as rationalization error.

Advantages :Usage of Random forest depicts the significance of non constantsineitherclassification or regression algorithms problemina prevalent way can be done by Random Forest. The 'amount feature is the transactional amount. Feature 'class' is the target class for the two types ofbinary classification and it takes 1 forpositive case this means fraud negative (0) means non fraud

Table: Raw features of credit card transactions

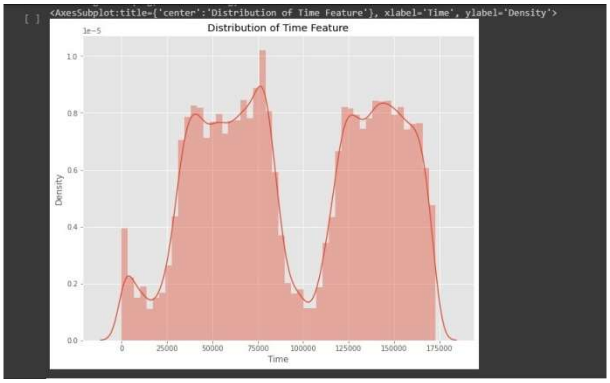
TABLEI. DATASET DESCRIPTION

Attribut e names	***Descriptio n***
Transacti o n id	Identification number of a transaction
Cardhold e r id	*Unique identification number given to the cardholder*
Amount	*Amount transferred or credited in a particular transaction by the customer*
Time	*Details like and date, to identify when the transaction was made*
Label	*To specify whether the transaction is fraudulent*

Details like and date, to identify when the transaction was made
Label
To specify whether the transaction is fraudulent

A. DATASETDESCRIPTION:

The dataset contains transactions made by a cardholderinatime during intwodaysthat is .,twodaysinamonth. In which the total 284,807 transactions are among in which there are 492that is ., 0.172% transactions are fraud and remaining are non fraudlenttransactions .



```
[ ]  <AxesSubplot:title={'center':'Distribution of Time Feature'}, xlabel='Time', ylabel='Density'>
```

Fig:1

This dataset seems to be highly imbalanced. Since using the data of daily transaction details of a customer is considered to the issue which relates to secrecy.

Therefore, majority of the columns names in the dataset are altered using principal component analysis that is PCA. V1, V2, V3, V4, V5, V6, V7..., V28 are the column names in PCA. Applied features and remaining i.e., ‘Time’, ‘Amount’, ‘class’ are Non-PCA column features , as shown in table2 below.

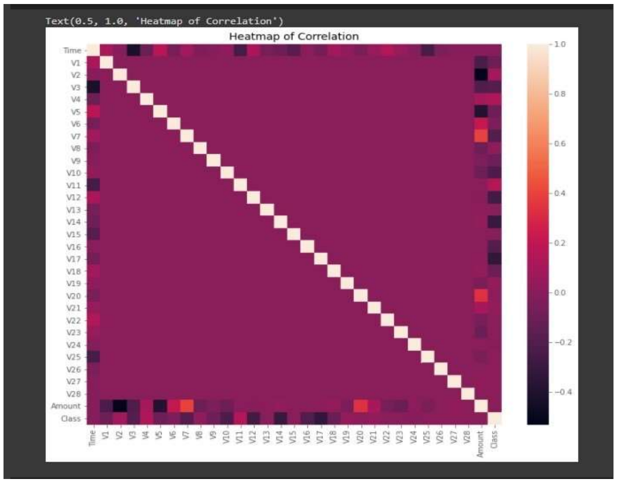
TABLEII. DESCRIPTIONOFFEATURESINDATASET

S. No	***features***	***Descriptio n***
1	*Time*	*Time in Second to indicate the elapses between n the current transaction and first transaction*
2	*Amount*	*Value of Transaction amount*
3	*Class*	*0-non fraud* *1-fraud*

a. Sample of a Table footnote. (Table footnote)

This Heatmap of Correlation explains that Class is independent of both the amount of transaction and time at which transaction was happend. It makes clear sense from the Heatmap of Correlation ,the class of the transactions is depends up on principal component analysis appliedattribute.

Show the correlation matrix of thedataset



```
Text(0.5, 1.0, 'Heatmap of Correlation')
```

Fig. 2.

Correlation Matrix for Attribute (both the X and Y axis show different attribute present in dataset.

I. CLASSIFICATIONALGORITHMS:

A. Logistic Regression:

Logistic Regression is classification based algorithms ,It is the Statistical model that in its basic form uses a logistic function to model a binary independent variable . For class problems due to its simplicity And effectiveness. It's far a statistical model used to classify The output into one or extra categories based on the Courting between the dependent and independent Variables.

B. Support Vector Machine:

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

To split the 2 classes of data points, there are many possible hyperplanes that might be chosen. Our objective is to discover a plane that has the maximum margin, i.E the most distance between data points of both lessons. Maximizing the margin distance gives a few reinforcement so that destiny records points can be categorized with greater self belief. .Support Vector Machine minimizes Overfitting by choosing a hyper With maximum margin of

Separation between the 2 instructions. With the aid of choosing the Appropriate kernel, price parameter

And gamma values, the efficiency of Support Vector Machine can be greatly increased.

C. Random Forest:

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Dissimilar. That is finished by means of constructing every tree Using separate bootstrapped examples of information and using reselected small dataset of the data attributes at each point of nodes

while building the unique trees Random forests Were proven to carry out better in comparison to svm And other strategies.

IV. PERFORMANCE METRICS:

A. *Area* under Curve - Receiver Operating Characteristics(AUC-ROC).The Area under Curve - Receiver Operating Characteristics is a performance metric graph for binary Classification issues. It Is a measure of the capability of the version to differentiate Between instructions. Higher the rating ,better the Overall performance.

The receiver operating characteristics (ROC) is drawn among TPR vs FPR

TPR = truepositive rate,

FPR = false positive rate ,

FN = false negatives,

TN= true negatives,

Fp = false positives,

where the TPP = TN/ (TP + FN) (1)TPR = TP/ (TP+FN) FPR = FP / (TN+FP)

This take account of to be a good metric for computing overall performance.

*F1Score*This gives the measure of the average of reciprocal of positive predicted value and sensitivity, Where

Positive predicted value = TPs/ (TP+Fp) Sensitivity = TP/ (TP + FNs)

F1 = (2 * Positive predicted value * sensitivityl)/ (Positive predicted value+ sensitivityl)

F-measure is Weighted arthimetic mean of both sensitivity And positive predicted value so increasing the score maximizes both the performance metrics.

It is more Accuracy especially when uneven class distribution

V. EXPERIMENT:

The dataset for this fraud detection problem has gathered From Kaggle [8] website ,The dataset is highly imbalanced with 0.172% being fraud cases and the rest legitimate that is non fraudulent Cases This Indicates The Fraud cases are very less that is highly Imbalanced.This Contains numeric values for input non-constants (i.e variables) those are results Of a principal component analysis conversion. Because of the privacy issues, the features which are original and more background info about the Data aren`t provided by the company. The main Problem here is dealing with imbalanced dataset for this under sampling the dataset randomly has effective for Imbalanced Dataset.Random under sampling is the majority class instances are discareded at random until a more balanced distribution is reached. A simple under-sampling technique is to under-sample the majority class randomly and uniformly. This can potentially lead to loss of information. But if the examples of the majority class are near to others, this method might yield good results.

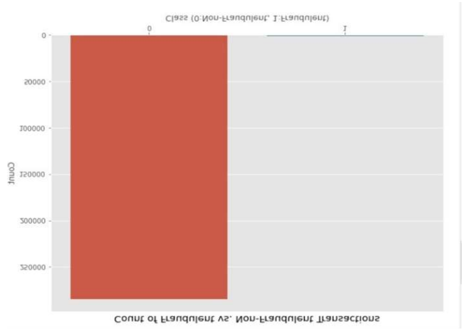


fig:3

Count of fraud vs transactions

Then we will discover the correlation among capabilities To pickout attributes with high positive and high poor associative Correlation. We dispose of the excessive weighted deviation (outliers) of the Dataset by using putting off samples that lie far away 2.5 multiple times of

IQR (this IQR is an estimate of deviation ,based on splitting the dataset as different quartiles). The Hyper parameters of aid vector device and random Forest are tuned by grid seek the use of curves as a scoring Metric. Grid search is process of scanning optimal parameters for a given model for every mixture

Of hyper parameters specified and evaluates every unique Model. The hyper parameters of the version with highest Score are then selected. A comparative analysis of the 3 classifiers is then Made with the chosen according to performance order to model the real- world scenarios where the Percentage of fraud type Transactions could be very small, we examine the Overall performance of the 3 classifiers that is random forest logistic regression and support vector machine with. One of a kind under sampling ratios – five%.,10% and 15%.

VI. RESULTS ANDDISCUSSION:

The consequences of the three diverse classifiers with Unique sampling ratios is proven within the below table underneath. As We are able to observe in below table ,and the ratio of fraudlent cases Decreases, the, the Area under Curve - Receiver Operating Characteristic curve And f1 rating of Suppport vector machine and logistic regression show a downward trend where as The ones of randomforest show better overall performance. Logistic Regression and random forests how some distance higher performances and contrast to support vector machines with respect to each overall Performance metric. The f-measure in binary classification of random forest is best as compared the other two with decreasing ratio of the fraudlent transactions. This Is an vital performance sign as it shows that the System is successfully classifying fraudlent transactions as well as minimizing errors in Incorrect classification ,both of which are extremely relevant to the real-worldscenario.

Table:

Performance over various database

Classifie r/ Under Sample ratios (Frau d class:Non Fraud)	***Logistic regression***		***Support Vector Machine***		***Random Forest***	
	AUC – ROC curve	***F-1 Score***	***AUC – ROC curve***	***F-1 Score***	***AUC – ROC curve***	***F-1 Score***
1:1	*0.971*	*0.923*	*0.928*	*0.917*	*0.944*	*0.930*
0:15:1	*0.972*	*0.941*	*0.966*	*0.926*	*0.958*	*0.937*
0.10:1	*0.972*	*0.944*	*0.970*	*0.926*	*0.967*	*0.940*
0.05:5	*0.972*	*0.924*	*0.940*	*0.921*	*0.966*	*0.940*

The overall performance of random forest is explored better If the various features of dataset are not anonymiysed and by Similarly tuning of the hyper parameters. If the capabilities of the Records are recognized, feature extraction may be performed to reduce The measurement of records. The tuning right here is accomplished by means of grid Search whereas in similarly have a look at, tuning can be stepped forward via The use of random search accompanied with the aid of grid seek to acquire Higher parameters.

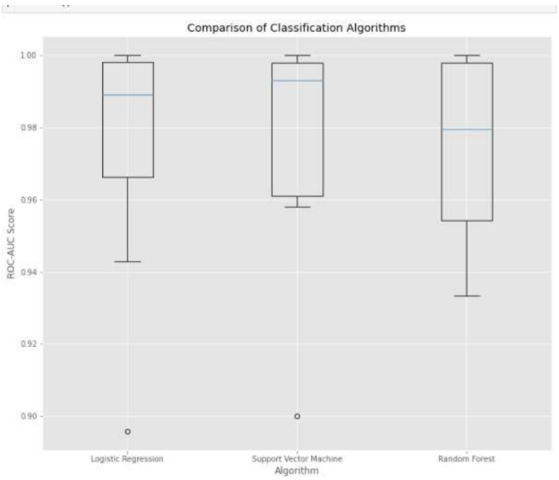


Fig. 4. Comparision of different algorithms

VIII CONCLUSION:

Thus the overall comparision For Credit Card Fraud with three classification algorithms(random forest ,support vector machine(SVM) and logistic Regression) by using under Sampling for imbalanced dataset with different percentage of dataset ratios that is 5% ,10% , 15%.random forest algorithm suited as best for dealing with imbalanced dataset by the evalution of Performance of each algorithm and also by plotting AUC-ROC curves,.The overall performance of random forest algorithm is further increased better by methods of Parameter tuning and feature extraction.

REFERENCES

[1] Learning Aowyemi, John O., al. "Credit Card Fraud Detection Using Machine Techniques: AComparativeAnalysis. In ML" 2017 International Conference on Computings Networking and Informatics (ICCNI), 2017,

[2] Mohammed, Emed, Dehrouz Far. "The Supervised Machine LearningAlgorithms for Credit CardFraudulentTransaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018.

[3] Randhwa, Kuldep, et al. "Credit Card Fraud Detection Using AdaBoost and Minority Voting." IEEEAccess,vol. 6, 2017, pp. 14277–14284.

[4] Gharsh, S., Reisly, D. L., "Credit card fraud detection with neural-network"s",In Sciences, Proceedings of theTwenty-Seventh Hawaii InternationalConference, vol. 3, pp. 621-630, 1994.

[5] Shrivastav, A., Kundun, A., Sutral, S., & Majumdar, A., "Credit card fraud detection using hidden Markovmodelmachine learning", IEEETransaction on dependableand securecomputings5(1),40-49, 2008.

[6] Xuang, S., Liu, G., Li, Z., Zheang, L., Wang, S., & Jiang, C., "Random forest for credit card fraud detection", IEEE 16th International Conference on Networking,Sensing and Control (ICNSC), vol. 1, pp. 1-5, 2016.

[7] Roy, Abhimantyu, et al. "Deep Learning Detection Fraud in Credit Card Transaction", Systems andInformations Engineering Designing Symposium (SIEDS), vol. 1, pp. 1-5, 2017.

[8] S. Amara and R. R. Subramanian, "Collaborating personalized recommender system and content-based recommender system using TextCorpus," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 105-109.

[9] Xuan, Shiyang, et al. "Random Forest for Credit Card FraudDetection.", IEEE 15th InternationalConference onNetworking,SensingandControl(ICNSC),2018.