Group 319 : Credit Card Fraud Detection Using Different types of Classification Techniques.

First Name	Last Name	Email address
Sarthak	Agrawal	sagrawal4@hawk.iit.edu
Shivani	Agrawal	sagrawal3@hawk.iit.edu

Table of Contents

1. Introduction

Credit cards are one of the most powerful sources provided by financial companies which gives freedom to cardholders to borrow funds for the purchase of materials. This borrowed amount must be paid by the borrower to the bank within certain days. With proper repayment of funds on time or due date builds a good credit score and hence helps in getting loans from banks. But few people misuse it by making fraud transactions .

Now the question arises how a credit card company detects the fraud transactions or whether the transaction is legal or not. so, the process starts from swapping of card which either accepts or rejects the transactions . But this is not the end , after your transaction is processed, they go through production models which helps in determining whether the transaction was fraud or legal . In this project our basic target is to find which classification model is used for prediction whether the predictive model is the best or are the other models with better predictions .

2. Data

The data set here used has been taken from Kaggle with almost 300,000 rows with 31 different attributes presented in the columns .

the data set has been taken it contains only numeric input variables which are the result of a PCA transformation. out of 31 columns 28 columns attribute is unknown for us which are often from PCA transformation whereas prime and amount columns are the transformations which have not been transformed and hence represents in the data set with their type.

Time:- which is described is the time in seconds between each transaction, where the type is numeric.

V 1 to V 28 :- are the features which are unknown to us as the data set obtained was using PCA transformation, hence they have removed the name of the features due to some confidentiality issues and the type is numeric.

Amount :- hear it is described as the money used in each transaction and the type is numeric.

Class :- It determines whether the transaction is fraud or legal and hence its type is Boolean.

Column	Description	Type
Time	in seconds taken between each transaction.	Numeric
V 1- V 28	unknown features due to confidentiality issues .	Numeric
Amount	money used for particular transaction	Numeric
Class	fraud or legal	Boolean

3. Problems and Solutions

When a card is copied or stolen the transaction made by them are labeled as fraud, they should be detected in a timely manner else results in Loss. to determine this bank uses two-layer detection first rule-based detection and statistical based detection.

The focus of our project is on statistical layer but it is not easy because amongst 10,000 transactions only few are detected as frauds hence the model which is being used now may not be effective in near future as the previously used techniques might fail we want new methods or techniques to be used therefore finding best predictive classification model is our aim using various plots, confusion matrix and finally using AUC, accuracy, precision, recall and ROC to determine which is the best model to use for the data set.

Potential solutions are different classification techniques using their confusion matrix through which we will predict ROC, AUC, accuracy and precision.

classification techniques used in this data set are:-

1. logistic regression
2. decision tree
3. random forest

4. KDD

4.1. Tools

The tools used here is Jupiter notebook and has been worked on R coding.

4.2. Data preprocessing

Step 1: Preprocessing of our data set starts with loading of data but before that we need to download some packages and load the various libraries that will be used in the data set or for the regression classification model we are going to use.

Step 2: Using read.CSV we have loaded our data set with the name credit card and checked whether there are some missing values in our dataset or not.

```
credit_card <- read.csv("creditcard.csv")

In [ ]:
creditcard <- credit_card

In [ ]:
apply(creditcard, 2, anyNA)   # checking if there is any NA
table(creditcard$Class)
```

Step 3: Setting up the seed and removing the time variable is the next step because time does not help us in any kind of predictions throughout our model, so we have removed it at first.

```
#----------setting the seed----------#
set.seed(4495)
creditcard$Time <- NULL ##### removing the time variable
creditcard[is.na(creditcard)] = -9999
```

Step 4: After checking if there are any values which are not available to us, we have removed them and replace them with the mean value of the dataset.

Step 5: Next is splitting the data into test and training dataset so that we can run various algorithms and predict.

```
#---------- creating partition ---------------#
set.seed(4495)
t<-createDataPartition(p=0.5,y=creditcard$Class,list = F)
training<-creditcard[t,]
testing<-creditcard[-t,]
```

Post Processing:

Step 6: Also creating synthetic data set and its graph as it is used to train the fraud detection system itself and in testing and creating other types of systems.

```
#--------------------- generating Synthetic data -------------------#
library(ROSE)
attach(training)
set.seed(4495)
training_Rose <- ROSE(Class~.,data=training,seed = 4495)$data
training_Rose$Class <- as.factor(training_Rose$Class)
ggplot(training_Rose,aes(x = Class)) + geom_bar(aes(fill = Class))
```

Step 7: For logistic regression classification model we need to find whether the data is balanced or imbalanced for that we must undergo two processes that is under sampling and oversampling

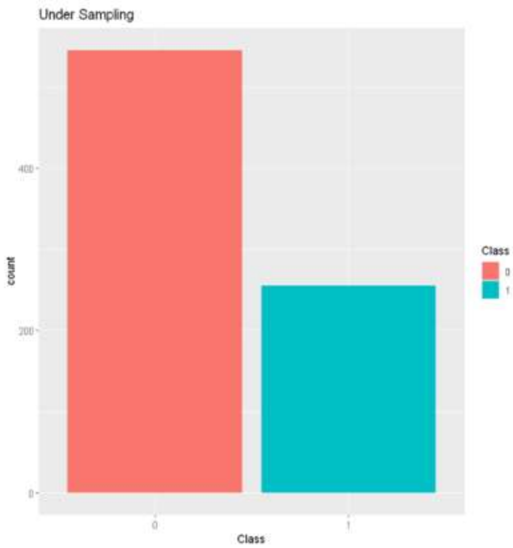
UNDERSAMPLING: It deletes or merges new synthetic examples in the majority class to exclude from the training data set and include selective samples from the majority class.

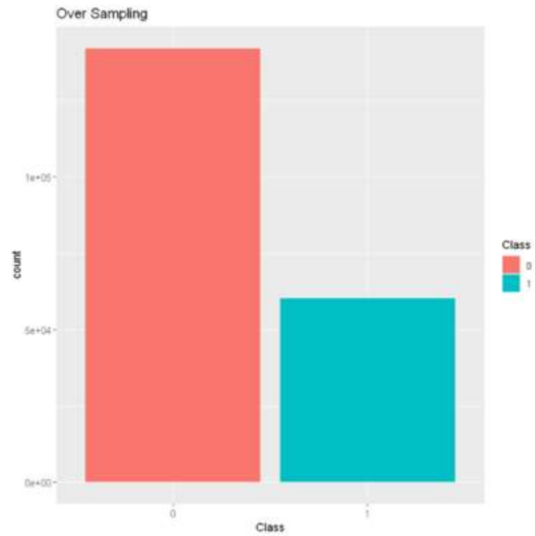
OVERSAMPLING: Creates or duplicates new synthetic example in the minority class to include from training dataset and exclude selective samples from minority class.

```
## ------------------- Undersampling -----------------------
training <- na.omit(training)
attach(training)
training$Class <- as.factor(training$Class)
training_under <- ovun.sample(Class~.,data = training,method = "under",
                              N=800,seed=4495)$data
ggplot(training_under,aes(x = Class)) + geom_bar(aes(fill = Class))+ggtitle("U
nder Sampling")
```

```
In [ ]:
## -------------------- oversampling -----------------------

training_over <- ovun.sample(Class~.,data = training,method = "over",
                             N=202404,seed=4495)$data
ggplot(training_over,aes(x = Class)) + geom_bar(aes(fill = Class))+ggtitle("Ov
er Sampling")
```





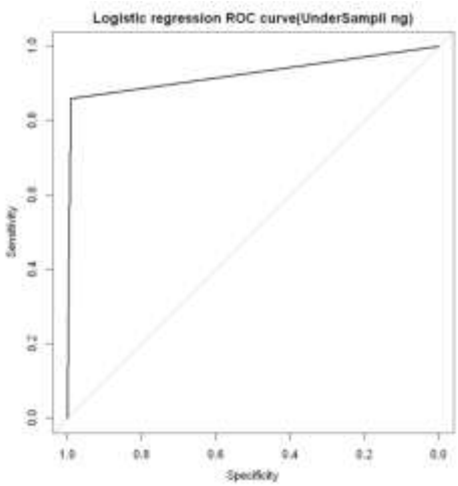
4.3. Data mining methods and processing

There are several techniques which can be used for the prediction but in this project, we have used only four techniques which are :-

1. **Logistic Regression**:- It is a classification algorithm that is used to predict a binary result based on a collection of independent variables as we deal with binary information in our project as our performance should be either yes or no, fraud or illegal. For credit card companies to issue a credit card, each person applying for it requires system or model to predict if the payments would default on a given customer. There are three types of logistic regressions.
   binary logistic regression multinomial and ordinal logistic regression. We will use binary logistic regression in our project.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

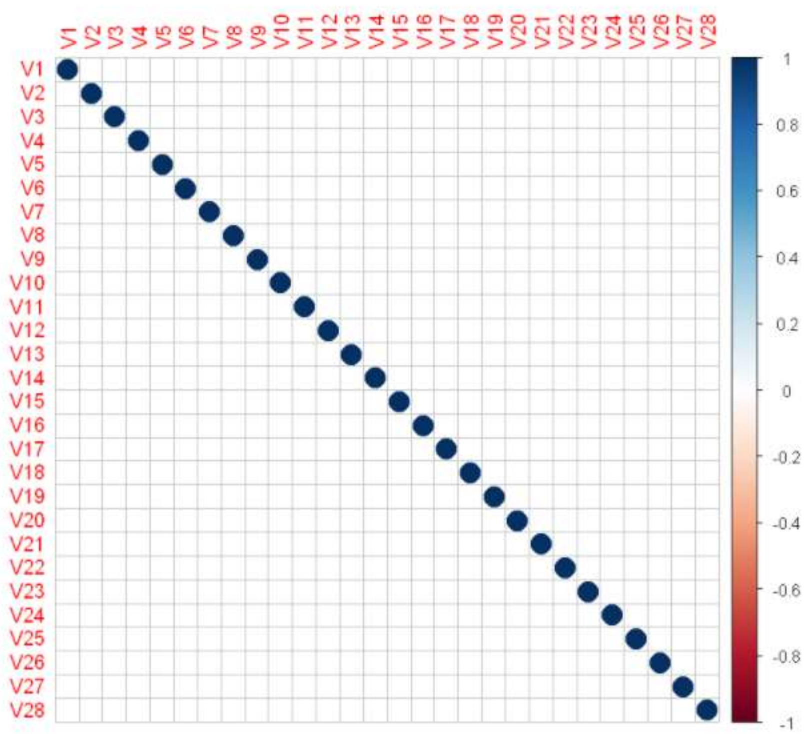
Results after Working on logistic regression classification method, AUROC curve, features importance and correlation matrix.



AUROC Curve

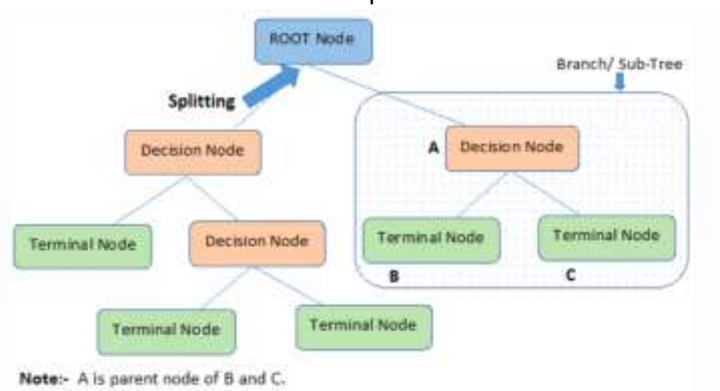
	<dbl>		
V1	20.7734055		
V2	9.1857692	V16	24.8075635
V3	16.7970361	V17	17.7684252
V4	55.7589186	V18	3.8484204
V5	20.1315767	V19	4.5958506
V6	12.3605734	V20	21.8016873
V7	12.1023691	V21	10.1083491
V8	18.5051715	V22	35.3937924
V9	15.2503371	V23	2.0585542
V10	30.2930195	V24	0.5703025
V11	24.5119709	V25	3.9919142
V12	34.2939813	V26	29.9084167
V13	30.5482485	V27	9.4307256
V14	41.9822865	V28	6.1087944
V15	3.3806752	Amount	17.7063166

Variable Importance

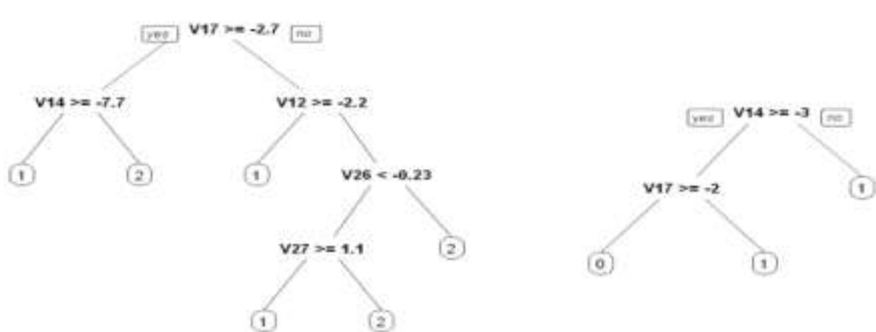


Correlation Matrix

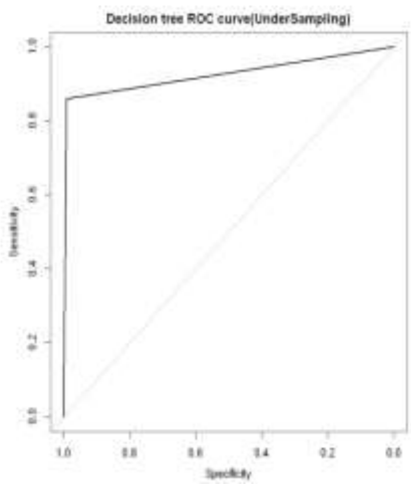
2. **<u>Decision tree</u>**:- Decision tree models are known as Classification Trees when the target variable uses a distinct set of values. Each node, or leaf, represents class labels in these trees, while the branches represent conjunctions of characteristics that lead to class labels. A decision tree where a constant value is taken by the target variable, usually numbers, is called Regression Trees. Feature selection is one of the most important components in decision trees it is based on what features of the data are relevant for the result we want to predict and decides the features accordingly.
   There are some important terminologies which are used in decision tree classification:-
   1. **Root node :-** it is the parent node which divides the data into two or more sets which is selected according to Attribute Selection Techniques.
   2. **Branch** :- any part of the decision tree.
   3. **Splitting** :- dividing the parent node into two or more child nodes using conditions if and else.
   4. **Decision node**:- dividing the child nodes into more sub-child nodes.
   5. **Terminal node**:- the last note which cannot be divide it further and hence is the end of the decision tree or can also be called as the final prediction.



<u>Results after Working on Decision Tree regression classification method, AUROC curve, features importance and final tree model.</u>



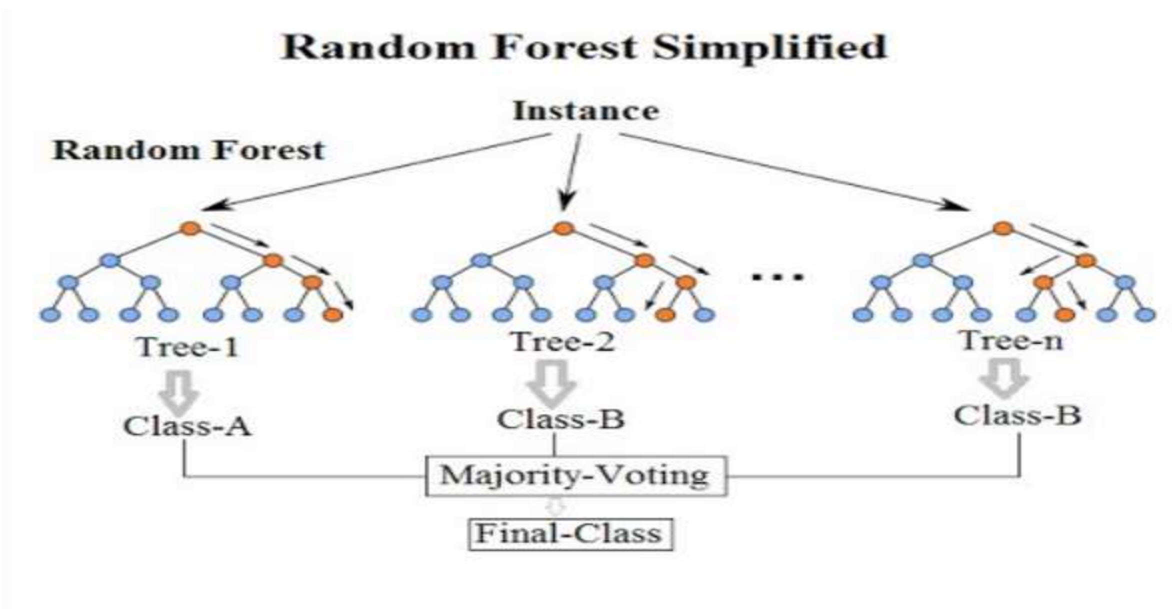
<u>Final Tree Model</u>



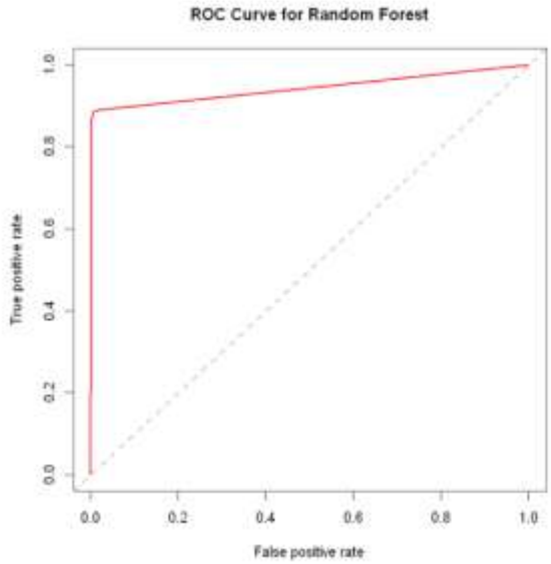
	<dbl>		
V10	64393.638		
V11	55568.386		
V12	56527.482	V9	0.000
V14	67253.315	V13	0.000
V17	60349.214	V15	0.000
V20	1889.982	V16	0.000
V4	2176.235	V18	0.000
V1	0.000	V19	0.000
V2	0.000	V21	0.000
V3	0.000	V22	0.000
V5	0.000	V23	0.000
V6	0.000	V24	0.000
V7	0.000	V25	0.000
V8	0.000	V26	0.000
V9	0.000	V27	0.000
V13	0.000	V28	0.000
V15	0.000	Amount	0.000

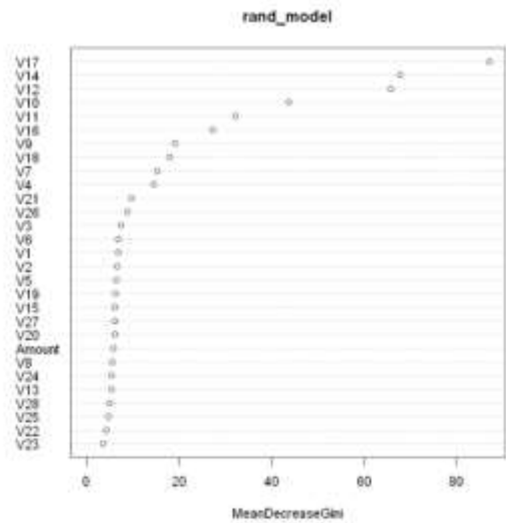
AUROC Curve   Variable Importance

3. **Random Forest Regression**:- Random forest plays a vital role in the classification process , it consists of a large number of decision trees which operate individually but every individual tree in the random forest classification model gives different predictions and the model with the most relevant outputs becomes our models prediction. Variable importance in random forest regression depends on the number of votes which has been casted for the correct class. The gini impurity criteria for the two descendant nodes is less than the parent node anytime a division of a node is made on variable. adding gini decrease overall trees in the forest for each individual variable which provides a simple variable significance that is also quite compatible with the measure of permutation value.



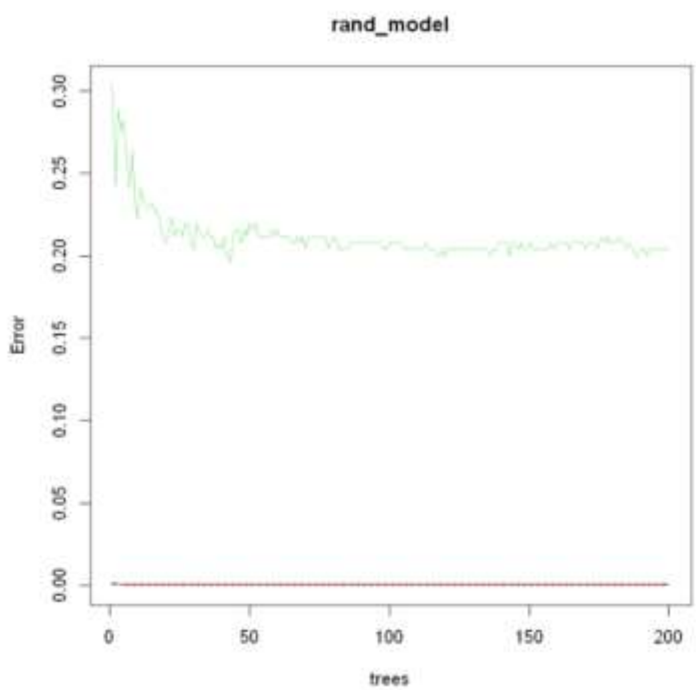
Results after Working on Random Forest regression classification method, AUROC curve, features importance and visualization graph.





AUROC Curve

Variable importance



Visualization Graph

5. Evaluations and Results

5.1. Evaluation Methods

there are multiple classification metrics used to determine which method is the best and gives the proper results. in our project we have basically worked on 4 metrics methods to evaluate which classification technique is the best and is most reliable for the proper predictions.

the classification matrix used are:-

	Predicted: NO	**Predicted: YES**
Actual: NO	TN	FP
Actual: YES	FN	TP

1. **Accuracy** :-

$$\text{Accuracy} = \text{fraction of correct classifications} = \frac{TP + TN}{TP + TN + FP + FN}$$

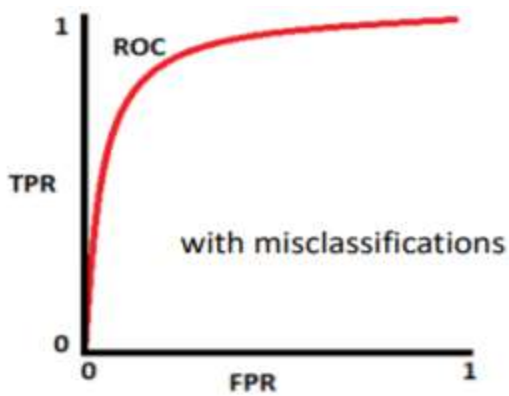
2. **Precision** :- also known as positive prediction value.

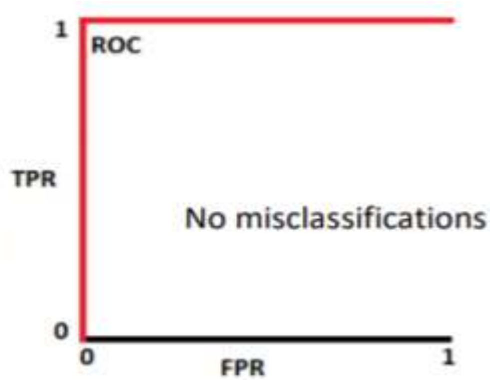
$$\text{Precision} = \begin{array}{c}\text{fraction of correct classifications} \\ \text{in the positively labeled results}\end{array} = \frac{TP}{TP + FP}$$

3. **Recall** :- also known as sensitivity.

$$\text{Recall} = \begin{array}{c}\text{fraction of positives} \\ \text{which were correctly predicted}\end{array} = \frac{TP}{TP + FN}$$

4. **AUROC** :- AUROC curve area Determines how accurate our predictions as larger the area under the curve (AUC) more accurate our predictions are.





$$\text{TPR} = \text{true positive rate} = \frac{TP}{TP + FN} \qquad \text{FPR} = \text{false positive rate} = \frac{FP}{FP + TN}$$

5.2. Results and Findings

After working with all the methods and classification used in our project, we have the results as:-

Methods used	Accuracy	Precision	Recall	AUROC
Logistic regression	99.07	86.07	99.09	0.925
Random Forest Regression	99.95	92.69	79.86	0.90
Decision Tree	99.22	83.12	99.25	0.925

6. Conclusions and Future Work

6.1. Conclusions

After working on all the methods use fault classification to predict which is the best model that we can use for the better credit card fraud detection method we have come to a conclusion that Decision tree classification method is the best amongst all with accuracy of 99.22% and AUROC of 0.925 .

6.2. Limitations

While using the above data set the issues generated worth some of the techniques couldn't be used due to a very larger data set.

While working on random forest regression model we faced a several issues while generating a proper confusion matrix , to avoid this problem I have calculated all the evaluation metrics on paper and then Replicated them in results in finding part.

Moreover, there were issues in creating feature/variable importance graphs, to overcome this problem we have not generated any crafts for feature importance's but we have created a table with generates and tells which particular feature has the most importance in that particular classification method used in prediction.

6.3. Potential Improvements or Future Work

In this project we have worked only on four techniques to determine the best predictive classification model. but for the future development or potential improvements we would like to work on few more models which are :-

1. GBM (gradient based algorithm)
2. XG boost
3. SVM (support vector machine)
4. Light GBM

7. References

1. https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53
2. www.google.com
3. www.kaggle.com
4. https://www.datacamp.com/community/tutorials/logistic-regression-R
5. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
6. https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/
7. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
8. Slides and Data from Blackboard by Prof. Yong Zheng.