**Machine Learning Strategy for Loan Prediction**

A DISSERTATION

Submitted to



Amity University, Kolkata

For the partial fulfilment of the award of the degree of

**BACHELOR OF SCIENCE IN STATISTICS (HONS.)**

By

SOUGATA KHAN

Enrolment No. A90555919014

Mail id- sougatakmp@gmail.com

Under the supervision of: **Prof. Debashish Roy**

Amity Institute of Applied Sciences

Amity University Kolkata

**Major Arterial Road (South East), Action Area II, Newtown**

**Kolkata – 700017**

(India)

**TABLE OF CONTENTS**

|  |
| --- |
|  |
| **ABSTRACT** |
| **CHAPTER ONE: INTRODUCTION** |
| **CHAPTER TWO: OBJECTIVE** |
| **CHAPTER THREE: DATA** |
| **CHAPTER FOUR: METHODOLOGY** |
| **CHAPTER FIVE: RESULTS** |
| **CHAPTER SIX: CONCLUSION** |
| **REFERENCES** |
|  |

Page: 1

Abstarct

Classification problem plays an important role in the machine learning domain. ML models such as Logistic Regression, K nearest Neighborhood, Decision Tree, Ensemble Learning like Random Forest, Boosting and Xgboost are commonly used to solve classification problems. Among this Algorithms Decision Tree, XGBoost, Logistic Regression are most used.

I have used here Logistic Regression and Random Forest for my project. In this case I have used data of a bank collected from kaggle. On the basis of the data we can predict that the individual may or may not get loan by building a suitable statistical model.

After that the data is split into two parts: Train and Testing. Then I build a logistic regression model and random forest model on Train Data and test the model under the test data. Comparing these two models I select one appropriate model with greater accuracy rate.

Introduction

The two most pressing issues in the banking sector are:

1) How risky is the borrower?

2) Should we lend to the borrower given the risk?

The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can

Page: 2

be used to assess a customer's loan status and build strategies. This model extracts and introduces the essential features of a borrower that influence the customer's loan status. Finally, it produces the planned performance (loan status). These reports make a bank manager's job simpler and quicker.

Objectives of the

The objective of this paper is to provide quick, immediate and easy way to choose the deserving applicants who wants loan from a bank among all applicants with the help of logistic regression and random forest classification.

DATA

I have the data of 252000 applicants who applied for loans. The data has six columns containing the headers:

1. Income (Annual income of the applicant)
2. Age ( Age of the applicant when he/she is applying for loan)
3. Experience ( Work experience of the applicant in years)How
4. Married / Single ( The applicant is married or single)
5. House Ownership ( Does the applicant own a house or he/she stays in rented house)
6. Car Ownership (Does the applicant own a car or not)
7. Profession ( What is the profession of the applicant)
8. City ( In which city the applicant stays)
9. State (In which state the applicant lives in)
10. Current Job Years (How long has he/she been doing this current job)
11. Current House Years ( How long has he/she been staying in the current house)
12. Risk Flag ( The applicant has been given the loan or not; i.e. Loan Given = 1 and loan not given = 0)

As we have income, so we don’t need “Profession” as a variable and as we have “House Ownership” & “Current House Years”, so we don’t need “City” & “State” as variable.

Page: 3

**Snapshot Of Data:**

Methodology of Project

When we have a dependent variable and one or more independent variables we may use Logistic Regression given the fact that our dependent variable has binary outcome. Normally we use Logistic regression to get the probability value or divide them into two categories like yes-no or 0-1. In simple linear regression, we modeled the mean my of the response variable y as a linear function of the explanatory variable: y = β0 +β1x ; But when y is just 1 or 0 (success or failure), the mean is the probability p of a success. Logistic regression models the mean p in terms of an explanatory variable x. We might try to relate p and x as in simple linear regression: p= µ = β0 +β1x. Unfortunately, this is not a good model. Whenever β1 ≠ 0, extreme values of x will give values of β0 +β1x that fall outside the range of possible values of p, 0 ≤ p ≤ 1. We use logistic regression because :

* The probability of the variable lies between 0 & 1
* Probability doesn’t vary linearlyx 1 {\displaystyle x\_{1}} x 2 {\displaystyl Y {\displaystyle Y} p = P ( Y = 1 ) {\displaystyle p=P(Y=1)}Y = 1 {\displaystyle Y=1}b {\displaystyle b} β i {\displaystyle \beta \_{i}}

This is is the equation of multiple linear regression:



Page: 4

Now, Let’s say instead of we are taking probabilities (P). But there is an issue here, the value of (P) will exceed 1 or go below 0 and we know that range of Probability is (0-1). To overcome this issue we take “odds” of P:

 { where p=probability of success}

Do you think we are done here? No, we are not. We know that odds can always be positive which means the range will always be (0,+∞ ). Odds are nothing but the ratio of the probability of success and probability of failure. Now the question comes out of so many other options to transform this why did we only take ‘odds’? Because odds are probably the easiest way to do this, that’s it.

The problem here is that the range is restricted and we don’t want a restricted range because if we do so then our correlation will decrease. By restricting the range we are actually decreasing the number of data points and of course, if we decrease our data points,

our correlation will decrease. It is difficult to model a variable that has a restricted range. To control this we take the log of odds which has a range from (-∞,+∞).



Now we just want a function of P because we want to predict probability right? not log of odds. To do so we will multiply by exponent on both sides and then solve for P.



Or, 

Or, 

This is the equation of logistics regression.

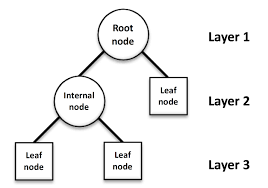
For logistic regression I used “dplyr”, “caTools”, “ROCR” library function and train the data. After that I test the model on test data and get confusion matrix but it is not so good, so I use another model Random Forest.

Page: 5

p = b β 0 + β 1 x 1 + β 2 x 2 b β 0 + β 1 x 1 + β 2 x 2 + 1 = 1 1 + b − ( β 0 + β 1 x 1 + β 2 x 2 ) {\displaystyle p={\frac {b^{\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}}{b^{\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}+1}}={\frac {1}{1+b^{-(\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2})}}}}

β i {\displaystyle \beta \_{i}} Y = 1 {\displaystyle Y=1} Y = 0 {\displaystyle Y=0} ( x 1 , x 2 ) {\displaystyle (x\_{1},x\_{2})} p {\displaystyle p} Y = 1 {\displaystyle Y=1} ….. Now, another classification problem is decision tree.b {\displaystyle b} The decision tree is tree shape diagram or chart that helps determine a course of action or show a statistical probability. The chart is called a decision tree due to its resemblance to the namesake plant, usually outlined as an upright or a horizontal diagram that branches out. Beginning from the decision it (called a "node"); each "branch" of decision tree which represents possible decision or outcome, or reaction. The furthest branches on the tree represent the end results of a certain decision pathway and are called the "leaves".

The decision tree is graphical depiction of a decision and every potential outcome or result of making that decision. Individuals deploy decision tree in variety of situations, from something simple and personal to more complex industrial, scientific, or microeconomic undertakings.

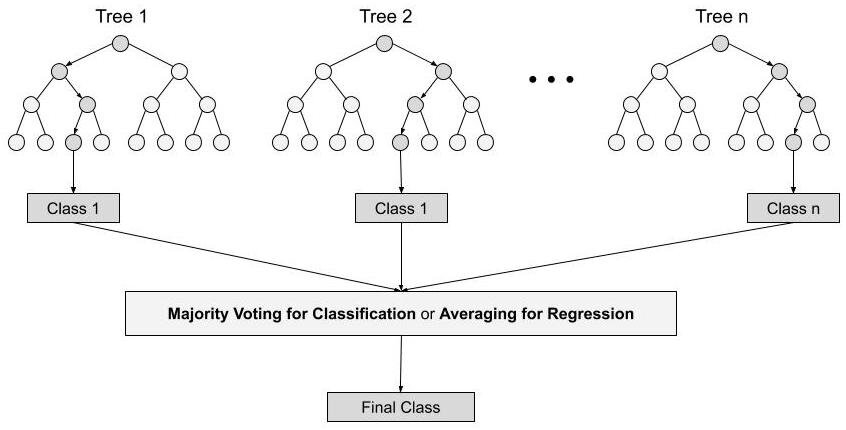


A random forest is a machine learning technique that’s used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

Bagging, also known as *Bootstrap Aggregation* is the ensemble technique used by random forest.Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as *row sampling*. This step of row sampling with replacement is called*bootstrap*. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as *aggregation*.

Page: 6



I have used logistic regression and random forest with R programming in this project. At first the dataset is imported in R-studio using read.csv. Here all variables have two or three factors like 0/1. I checked the data structure after that and notice that some variables are factor and some are integer. So I convert all the variables to factors.

Then I check the percentage of loan not given in dataset is more than loan given case and I need such a model so that its confusion matrix accuracy rate is higher than the percentage of loan not given case. So, I fixed this data disbalance using the weights of the dataset.

Then I splitted the data into two parts : train data set and test data set. Then I realize that I will use logistic regression for building a model since here all variables are factor and this is a classification problem.

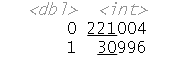
But my data is unbalanced so I use another classification model named Random Forest and this algorithm will give us a model with highest accuracy rate.

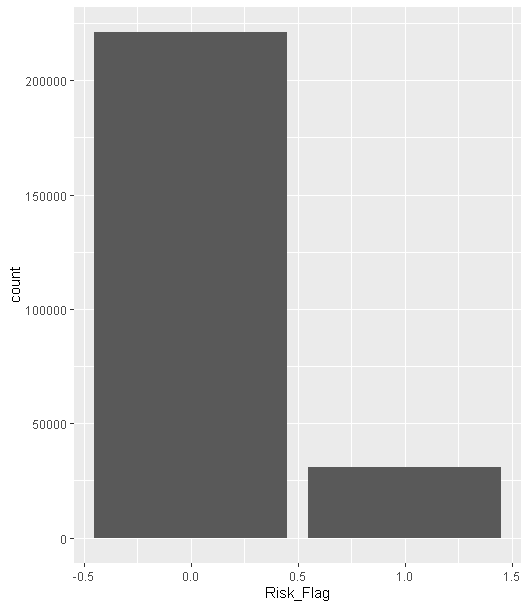
Before using the models I split the dataset in 75:25 ratios that means 75% of dataset is train data by which I can build the model and 25% of dataset is test dataset by which I can test my model accuracy.

Page: 7

For Random Forest algorithm, I use “randomForest”, “caret” library functions. In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. Here total no. of trees are 200. After that I find the confusion matrix and calculate the accuracy rate of model on test dataset. At last I save my Random forest model and I choose Random Forest as best model for this project.

Results

At first I find that there is 221004 no. of loan not given and 30996 no. of loan given case in my dataset.

Graph:

So I need a model that gives me a confusion matrix (as this is a classification problem) with greater than 71.30% accuracy.

To fix this data disbalance I gave weights to the “risk flag” column and wrote the following code to fix data disbalance.

Page: 8

“

w\_fact <- sum(train\_data$Risk\_Flag)/sum(1-train\_data$Risk\_Flag)

train\_data[c("Weights")] <- (train\_data$Risk\_Flag == 0) \* w\_fact + (train\_data$Risk\_Flag == 1) \* 1

“

This code fixed the data disbalance case.

Then the dataset is splitted in 2 parts : 75% for train and 25% for test , by using this code :

“

set.seed(188)

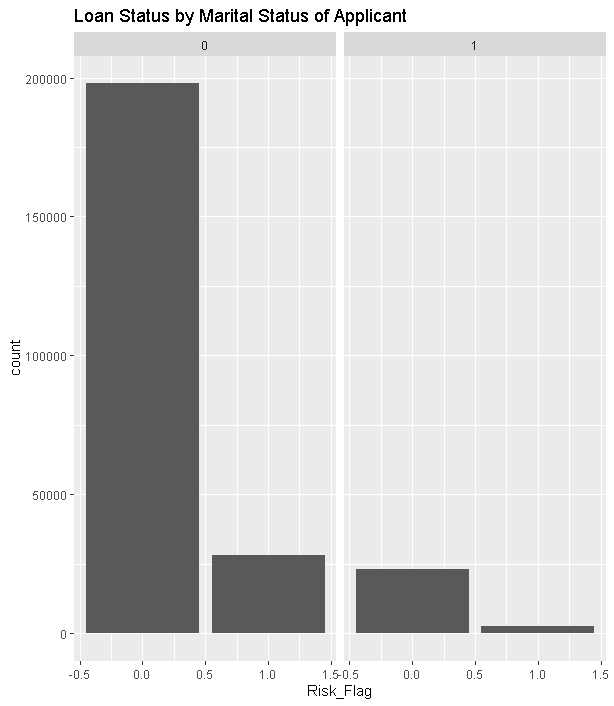
split=sample.split(train\_data$Risk\_Flag,SplitRatio= 0.75)

train\_train\_set=subset(train\_data,split==TRUE)

train\_test\_set= subset(train\_data,split==FALSE)

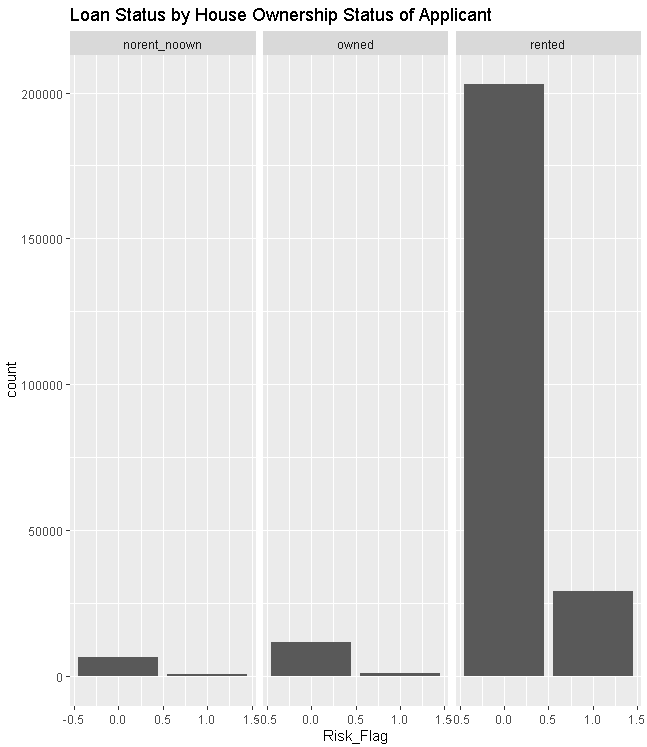
“

Then I did some exploratory data analysis :



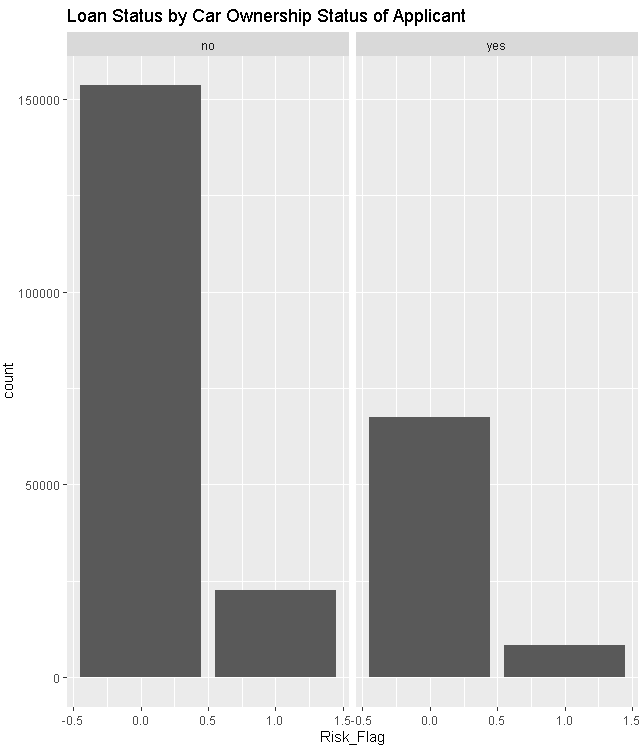
Page: 9

We can say that unmarried people are getting loan easily.



Page: 10

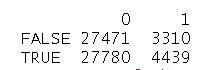
Visually someone can say that more people has got loan though they have rented house, but in comparison between a group a person has got loan easily if he/she has own house.



Page: 11

In this also visually someone can say that more people has got loan though they have no car, but in comparison between a group a person has got loan easily if he/she owns a car.

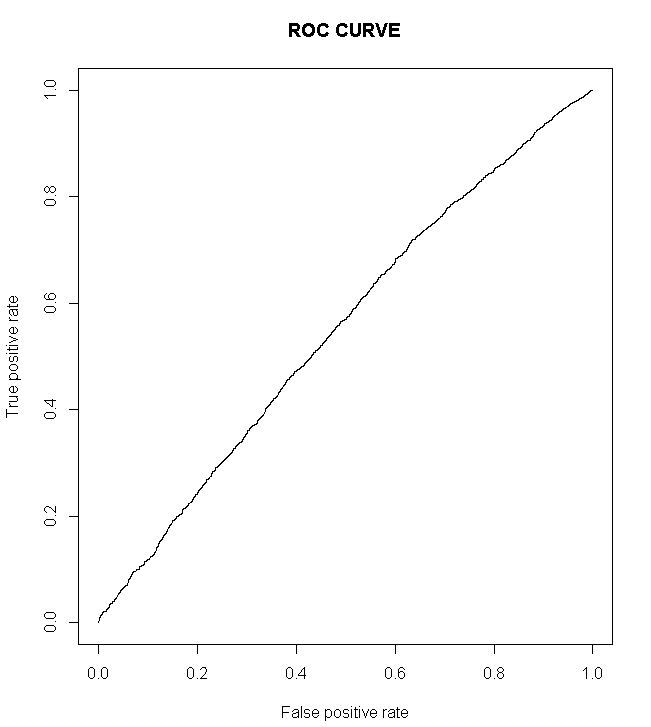
After getting the logistic regression model, dependent variable is predicted and calculated the confusion matrix for test dataset the confusion matrix of logistics regression algorithm.



Here accuracy rate is 50.65%. Formula of accuracy rate =

(True positive + true negative)/ (total observation).

Here the ROC curve is like :

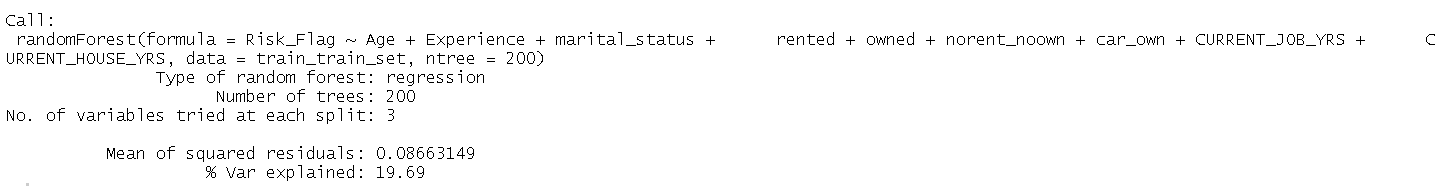


Page: 12

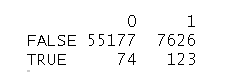
The accuracy of ROC curve is 87.71%. And area under ROC curve is 55.01%

After that I think I fit another algorithm on this data and comparing this two I will decide which one I will choose. So I choose Random Forest algorithm since it has so many advantages which I discuss before.

After converting, I run This random forest model on the train data with number of decision tree is 200. Then random forest model gave us the the following outcome:

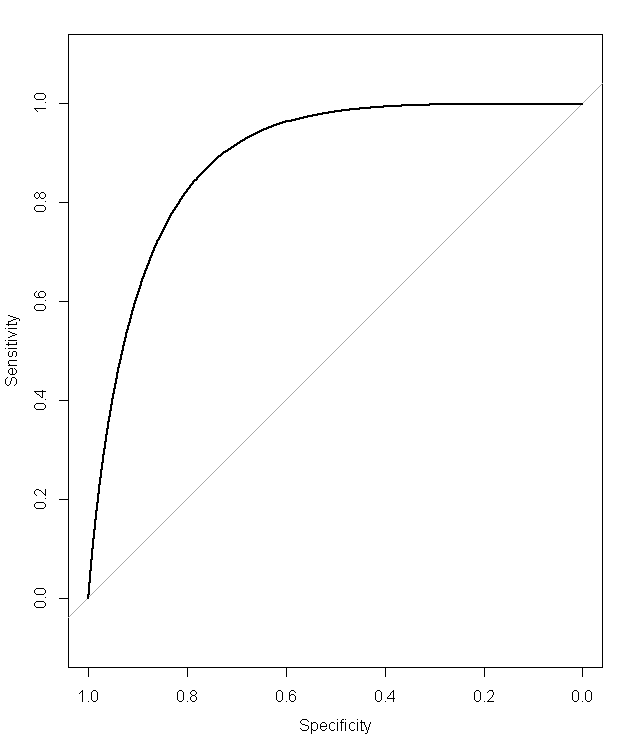


Now the confusion matrix of test data of random forest model is:



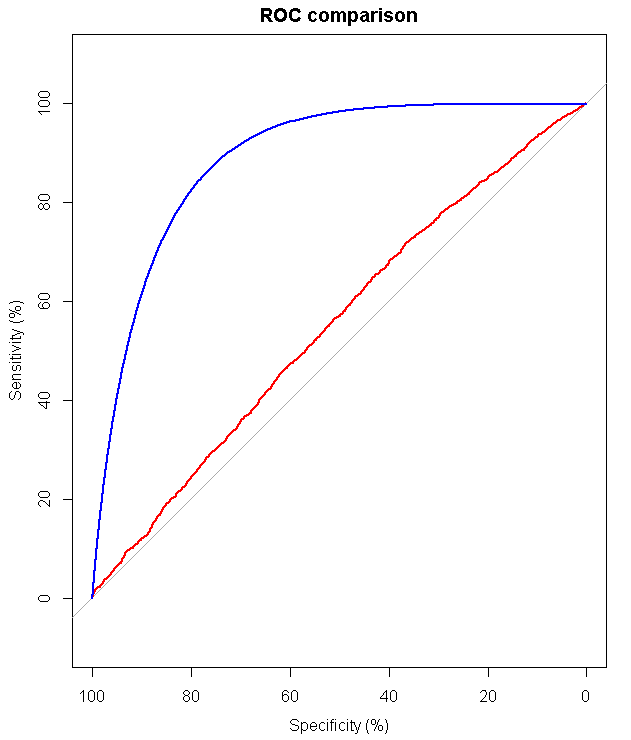
Now the accuracy rate of the confusion matrix is 87.77% and the ROC curve is like :

Page: 13



The area under ROC curve is 89.43%

So, if I compare the ROC curves of logistics regression and random forest model then it will clear that random forest has a better fit.



Page: 14

The blue line is for Random Forest algorithm and red line is for Logistics Regression algorithm.

Conclusion

|  |
| --- |
|  |

I have studied Logistics Regression and Random Forest. I did this work with both models. But the logistics regression model is not so good. Logistics Regression’s confusion matrix accuracy rate is not so high. Secondly, the ROC curve for Logistics Regression is not so good. The area under curve of ROC curve of Logistics Regression is not so high. Besides, the Random Forest model is better. Because its confusion matrix’s accuracy rate is higher than logistics regression’s confusion matrix accuracy rate. The area under ROC curve of Random Forest model is higher than that of Logistics regression’s.

Page: 15

So Random Forest model is chosen as a best fitting model for my project.

**Limitations :** There are another classification models like Decision Tree, XgBoost, K nearest Neighborhood etc. A better model may be made from those models.

Reference

Data: [www.kaggle.com](http://www.kaggle.com)

Logistics Regression algorithm : <https://en.wikipedia.org/wiki/Logistic_regression>

Logistics Regression in R : <https://www.tutorialspoint.com/r/r_logistic_regression.htm>

Random Forest in R : <https://www.tutorialspoint.com/r/r_random_forest.htm>

Books : 1) “Logistic Regression Models” by Joseph M. Hilbe

2) “Applied logistic regression” by David W. Hosmer

3) “Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis” by Frank E. Harrell, Jr.