PREDICTION OF CUSTOMER CHURN

RAHUL BHATTACHARYYA

22nd June 2018

Contents

1 Chapter 1
Introduction

1.1 Problem Statement

An organization loses its customers to its competition for various reasons. This particular case is related to telecom industry where a particular organization want to know given certain parameters whether a person will churn out or not. Mainly a customer churns due to new offers from the new company. After sales customer care can also be a reason for the change. However, if the potential target customers can be identified, then effective communication will surely be effective to reduce the churn percentage. In this problem we would like to predict whether organization will lose a particular customer to its competition or not.

1.2 Data

The main task is to build a classification model which will tell whether a particular customer will churn out or not. Below is the glimpse of the data

Column (1-6)

state	account length	area code	phone number	international plan	voice mail plan
HI	101	510	354-8815	no	no
MT	137	510	381-7211	no	no
OH	103	408	411-9481	no	yes
NM	99	415	418-9100	no	no
SC	108	415	413-3643	no	no
IA	117	415	375-6180	no	no

Column (7-12)

number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls
0	70.9	123	12.05	211.9	73
0	223.6	86	38.01	244.8	139
29	294.7	95	50.1	237.3	105
0	216.8	123	36.86	126.4	88
0	197.4	78	33.56	124	101
0	226.5	85	38.51	141.6	68

Column (13-18)

total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls
18.01	236	73	10.62	10.6	3
20.81	94.2	81	4.24	9.5	7
20.17	300.3	127	13.51	13.7	6
10.74	220.6	82	9.93	15.7	2
10.54	204.5	107	9.2	7.7	4
12.04	223	90	10.04	6.9	5

Column (19-21)

total intl charge	number customer service calls	Churn
2.86	3	False.
2.57	0	False.
3.7	1	False.
4.24	1	False.
2.08	2	False.
1.86	1	False.

So, in total we have 20 variables from which we would like to predict churn variable. The predictor variables as per the problems are.

1. account length
2. international plan
3. voicemail plan
4. number of voicemail messages
5. total day minutes used
6. day calls made
7. total day charge
8. total evening minutes
9. total evening calls
10. total evening charge
11. total night minutes
12. total night calls
13. total night charge
14. total international minutes used
15. total international calls made
16. total international charge
17. number of customer service calls made

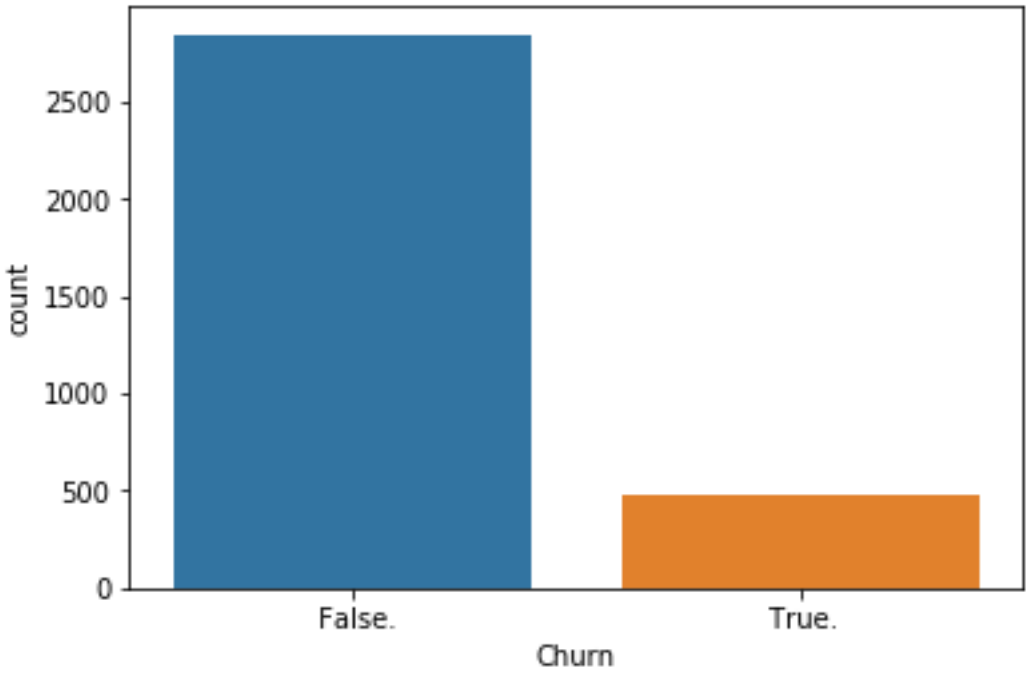
2 Chapter 2
Methodology

2.1 Pre Processing

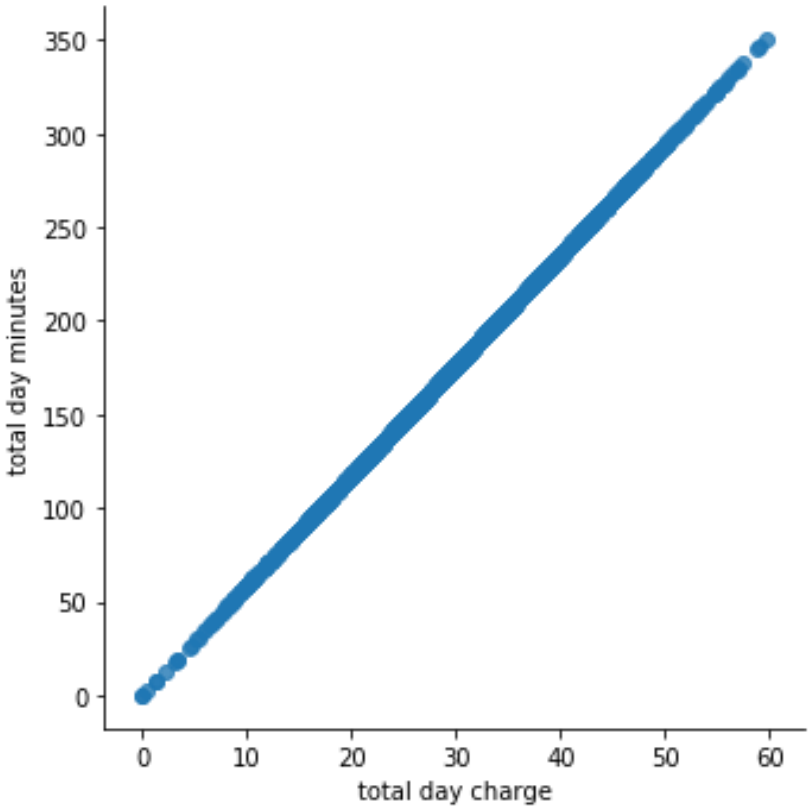
Any machine learning algorithms needs preprocessing techniques to prepare the data efficiently to fit the algorithms. In this problem several preprocessing methods like exploratory data analysis, chi-square test, Anova analysis, scaling techniques are applied to select and prepare the actual data. It was also noticed there was no missing value in the entire data. The outlier analysis is being avoided for this particular dataset as an outlier in the numerical variables can have valuable information which can be really useful from the model perspective. Example: If there is an outlier in total day calls made, an outlier which essentially a very low or high value can be meaningful. Also before proceeding with exploratory analysis "state", "area code", "phone number" have been dropped.

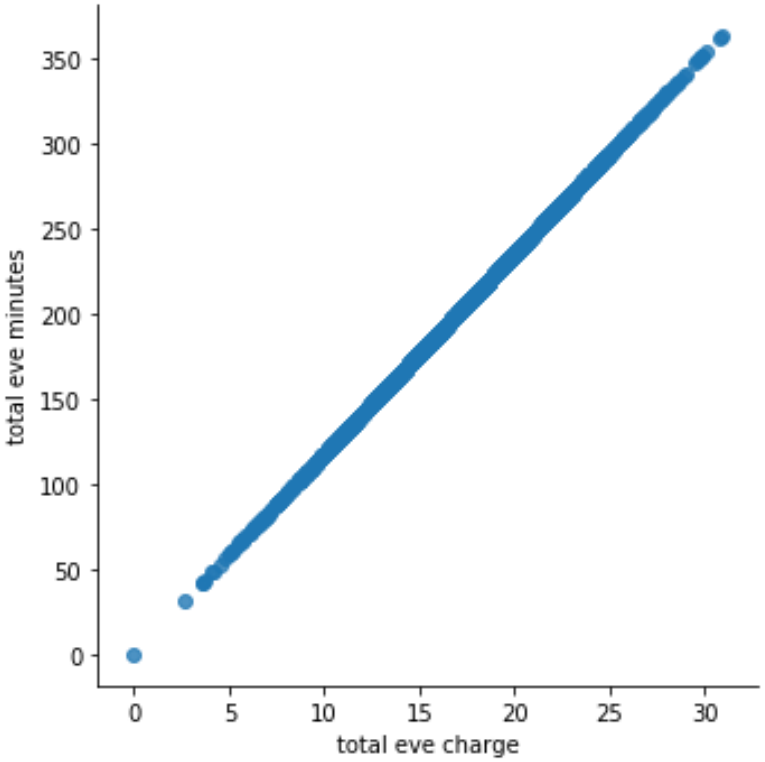
2.1.1 Exploratory analysis

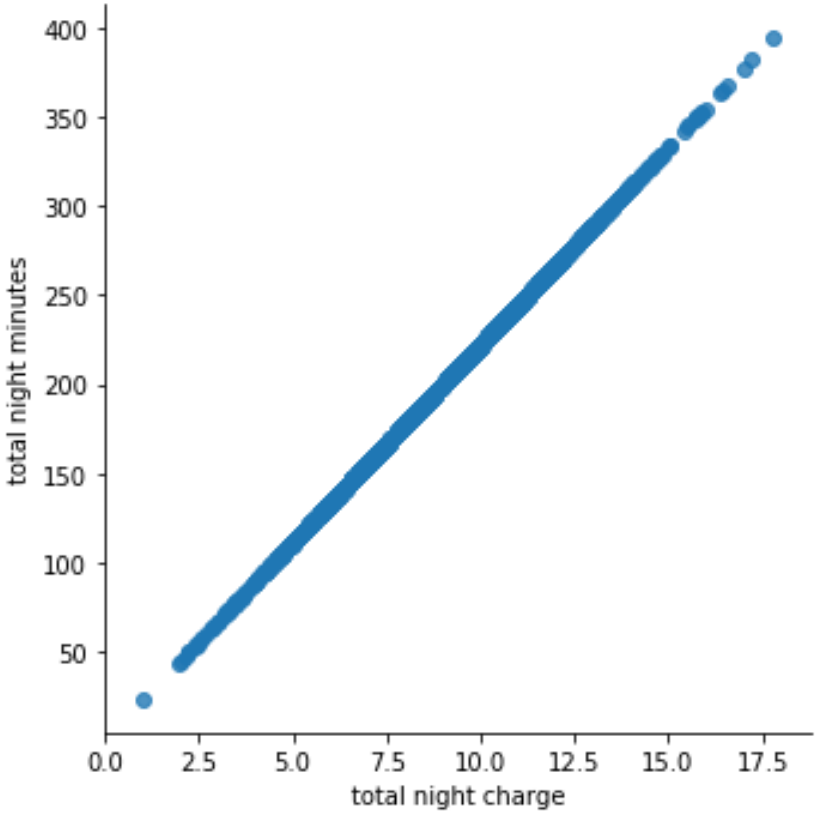
First, we have done an exploratory analysis on the target variable which is Churn. Our objective is to see the underlying distribution of two class and we can clearly see that our target data is highly imbalanced. The occurrence of False is way higher than True.

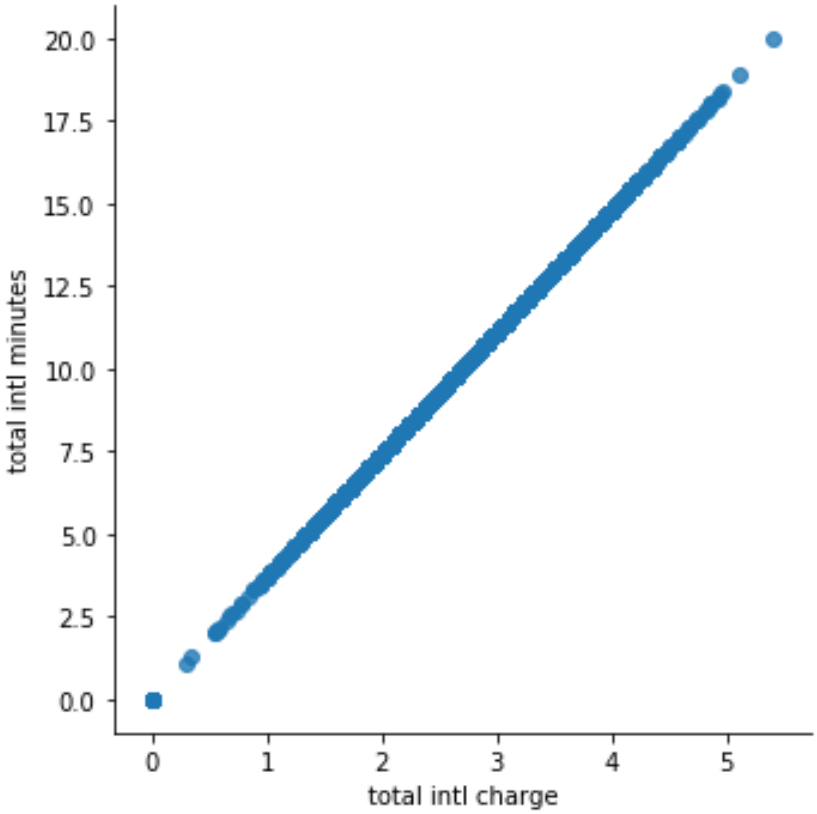


It can also be seen intuitively that total minutes used and total day minutes should be related to the total charge. so scatterplot of these types of variables are plotted and it confirms that they are linearly related. The variables used are 1) total day minutes used and total day charge. 2) total evening minutes and total evening charge 3) total night minutes and total night charge 4) total international minutes used and total international charge.


total day minutes
total day charge
0
50
100
150
200
250
300
350
0
10
20
30
40
50
60


total eve minutes
total eve charge
0
50
100
150
200
250
300
350
0
5
10
15
20
25
30


total night minutes
400
350
300
250
200
150
100
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
total night charge


total intl minutes
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
5
total intl charge

2.1.2 Feature extraction

The insight gained from the exploratory analysis tells us use of all the variables are unnecessary. Instead a ratio of those variables are used namely charge/minutes used for all four day, night, evening and international. so our new four predictor variables are

1. day_charge/minute

2. eve_charge/minute

3. night_charge/minute

4. intl_charge/minute

and the following variables are discarded.

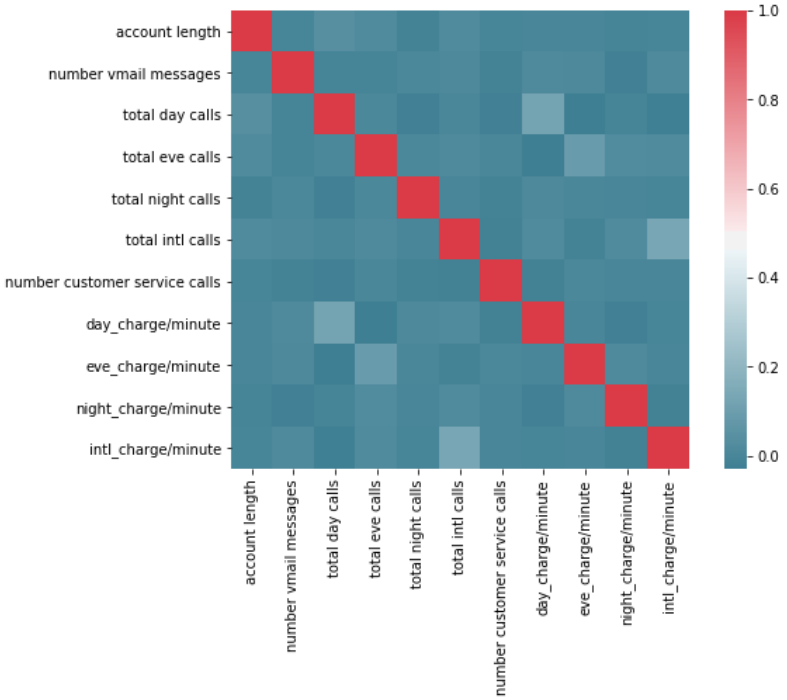
1. total day minutes used
2. total day charge
3. total evening minutes
4. total evening charge
5. total night minutes
6. total night charge
7. total international minutes used
8. total international charge

2.1.3 Feature selection

Before proceeding with this part the string values of "voicemail plan" and "international plan" are converted into numerical (from 'yes', 'no' to 1,0 respectively). The "Churn" variable also treated same way (from 'False.', 'True.' to 0,1 respectively). All these predictor variables are categorical and they are converted into that. A chi-square test is employed with these variables and 'Churn' variable to understand which variables are actually important for prediction purpose. The selection is based on the p-value produced from the chi-square test and no variable is discarded based on this test. In case of 'Churn' variable the 1 and 0 needs to be converted into 'yes' and 'no' respectively for python. However, R can internally take care of this.

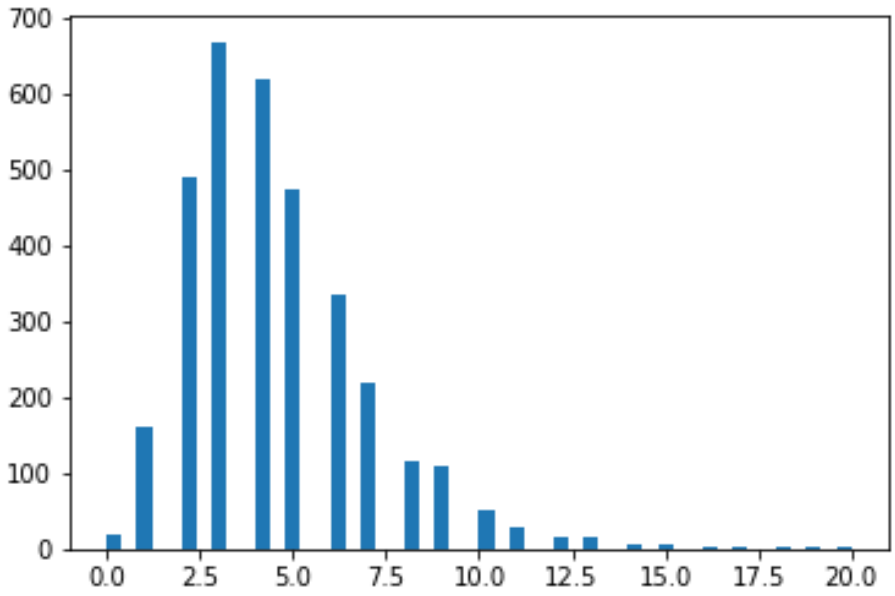
A similar anova test also employed for the numerical variables and no variables are discarded in this case also.

Also a correlation analysis shows that no variables are related to each other strongly. It can be seen from the following figure.



2.1.4 Scaling techniques

As some of the ratio variables like day_charge/minute, eve_charge/minute are between 0 and 1 and some variables like "account length" varies from 1-243 so we decided to scale the numerical variables. As both normalization and standardization techniques are available the distribution of the variables are explored. It has been found some of the features are having a skewed distribution which can be seen from the following histogram. So we employed normalization as scaling technique.



2.1.5 Oversampling technique

It can be seen from the exploratory analysis the target class is highly imbalanced. It can affect some machine learning algorithm in such a way that that the algorithm is biased towards a certain group. In order to eliminate this problem, we employed SMOTE algorithm to oversample the minority class of the target variable.

2.2 Modelling

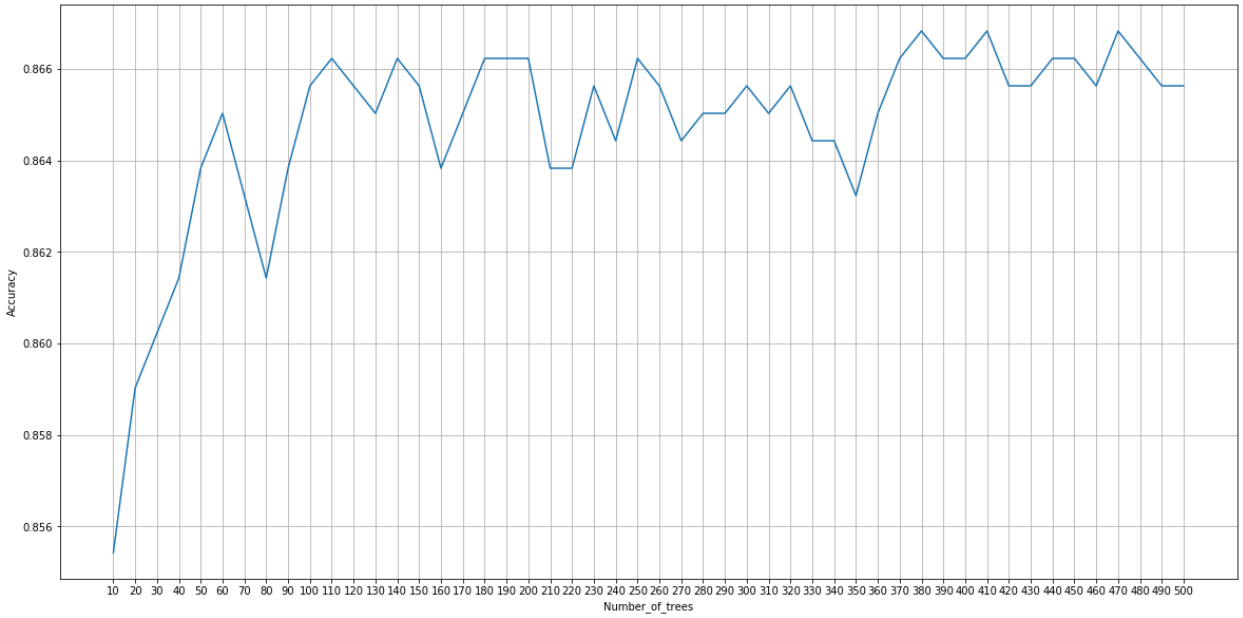
As this problem is a binary classification problem so we have employed several algorithms to predict our target variable. The models are:

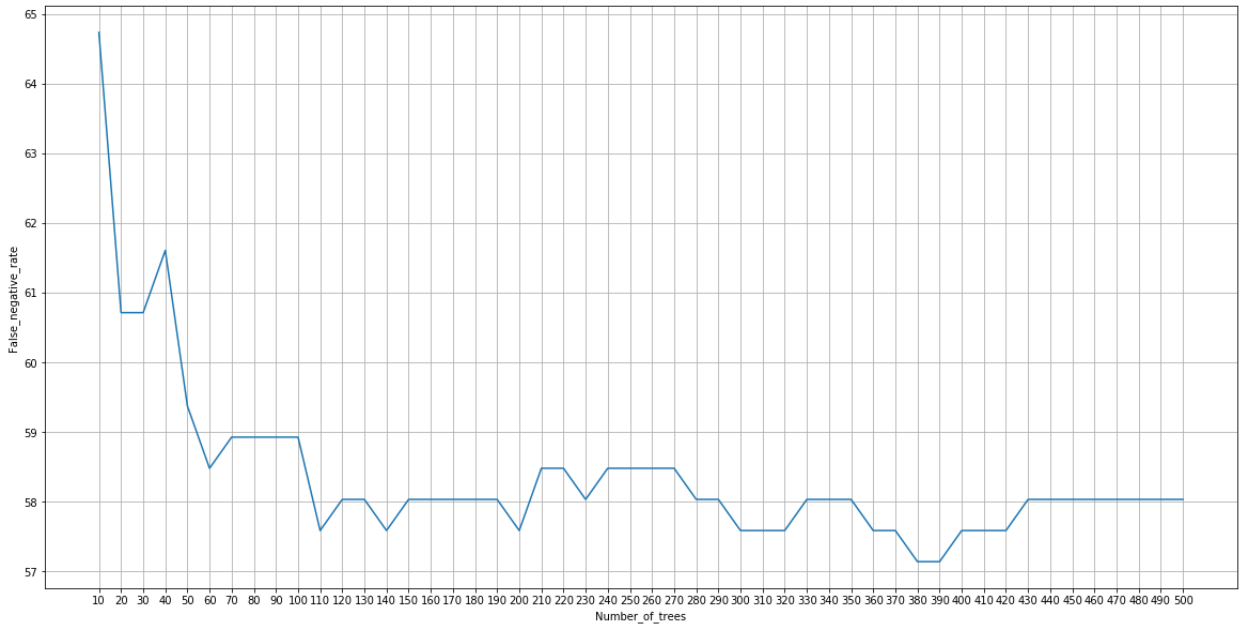
a) Random forest algorithm
b) Decision tree classifier (entropy criterion)
c) Logistic regression
d) Naïve Bayes classier (Gaussian)
e) Support Vector classification
f) KNN classifier

In this context false negative rate is chosen as error metric along with the accuracy. As for this particular problem customer wants to know the potential customer who may churn out, a high negative rate is highly regarded as negative because company may lose its customer.

In this case 'Train_data.csv' is used as our training purpose and 'Test_data.csv' for validation purpose. However train_test_split() method can also be used on train data to make a stratified sampling on a suitable ratio and use those two parts as train and test data. In case of R createDataPartition() method from caret package can be used to do the same.

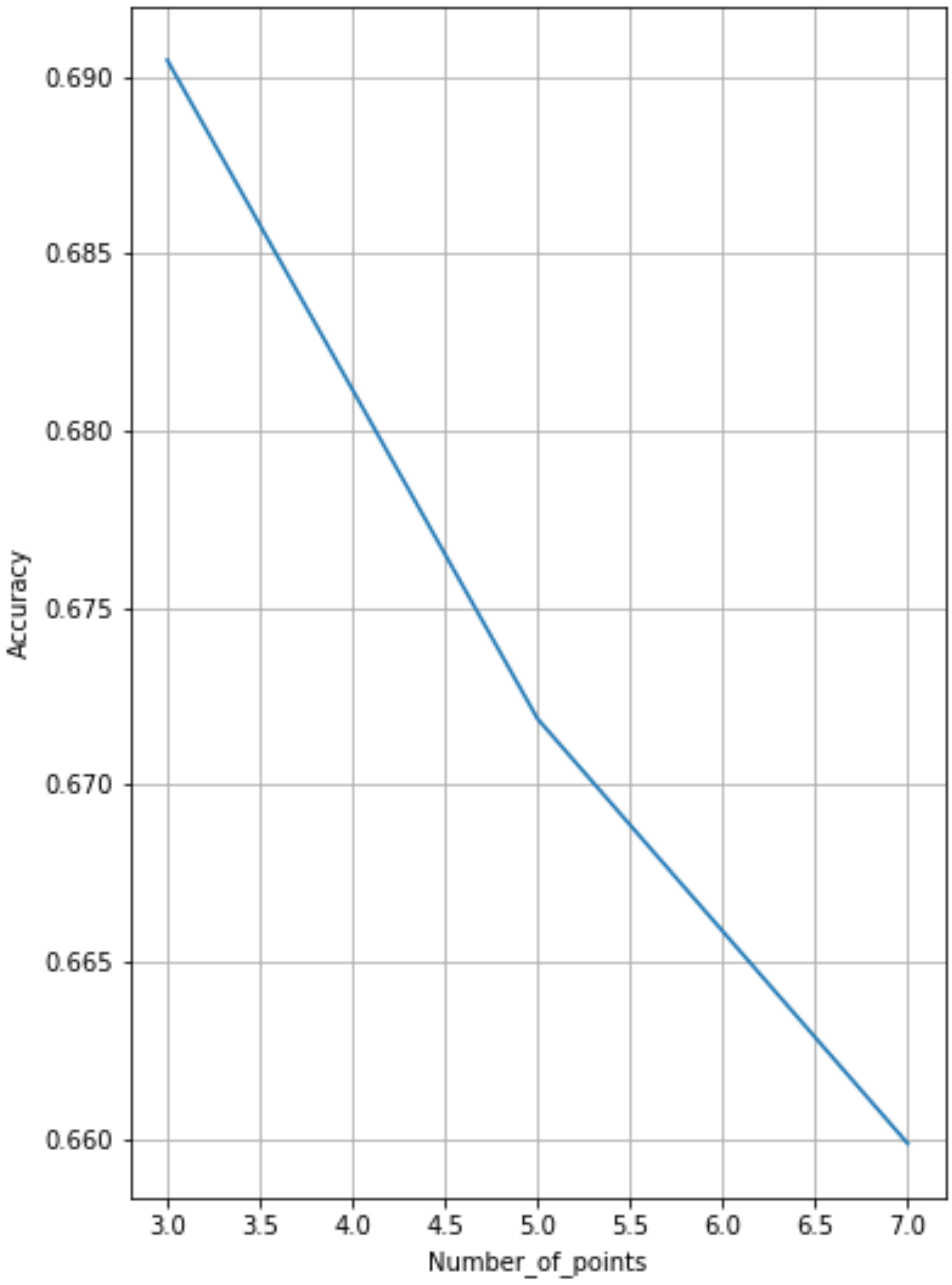
In case of random forest algorithm, we employed an iterative algorithm to choose correct number of decision trees. The algorithm iteratively calculates accuracy and false negative rate and plot it in a graph to choose the number of trees in an optimized way.

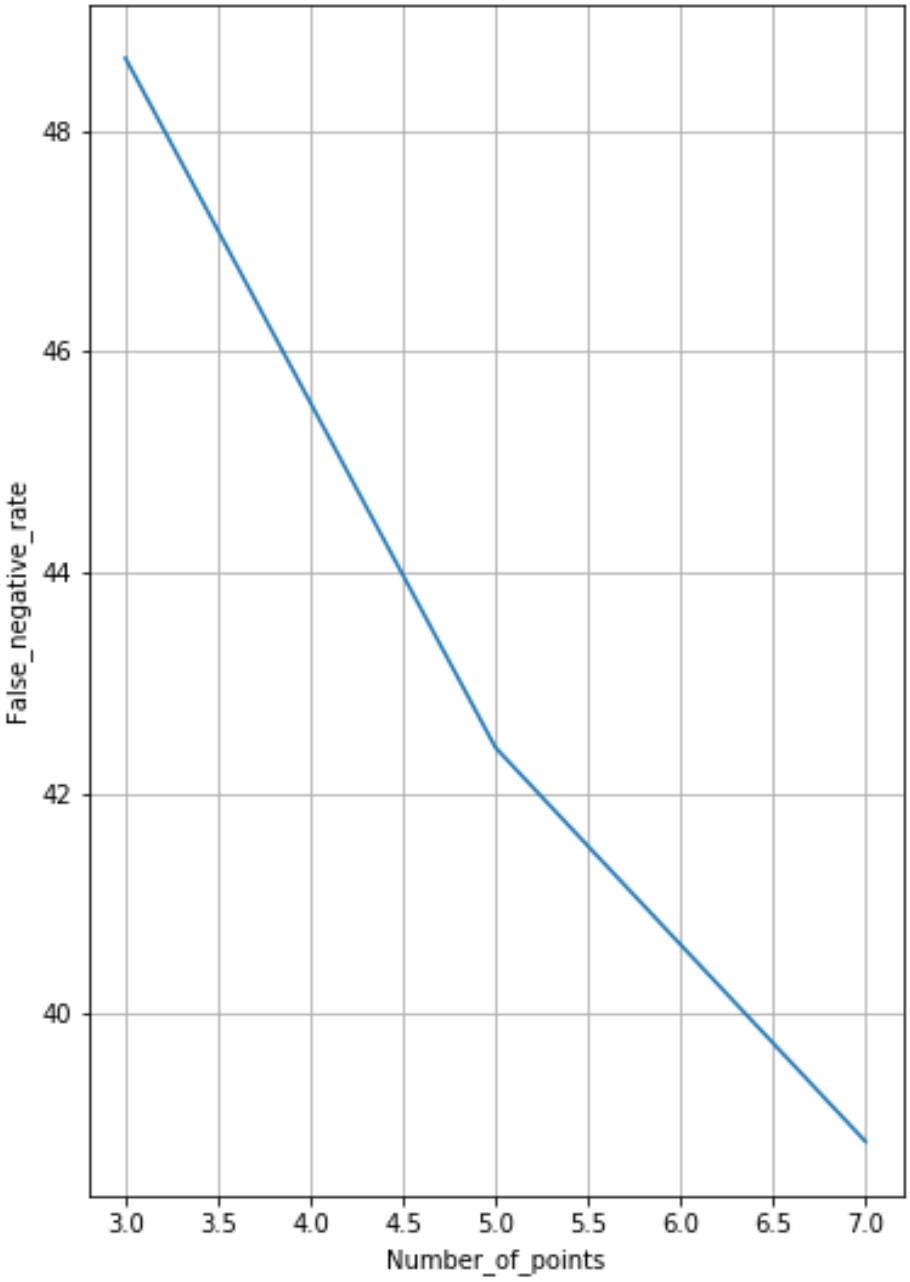




It can be noticed 380 is our most optimized no of trees to proceed with as it gives highest accuracy along with lowest False negative rate. The number of trees chosen in R is 50. The supporting figures can be found in the Appendix.

The decision tree classifier method is used with maximum depth of 10 (maximum path from tree to leaf). Logistic regression also employed. The Naïve Bayes classifier employed with a Gaussian assumption. Support vector classifier is employed with GridSearchCV() method to tune the parameters correctly. It has been found the best parameters are kernel=rbf, gamma=1, C=1000. Also KNN classifier method with number of neighbors 3,5 and 7 are used. It has been found 7 gives us the most optimized result, which can be seen from the following figures

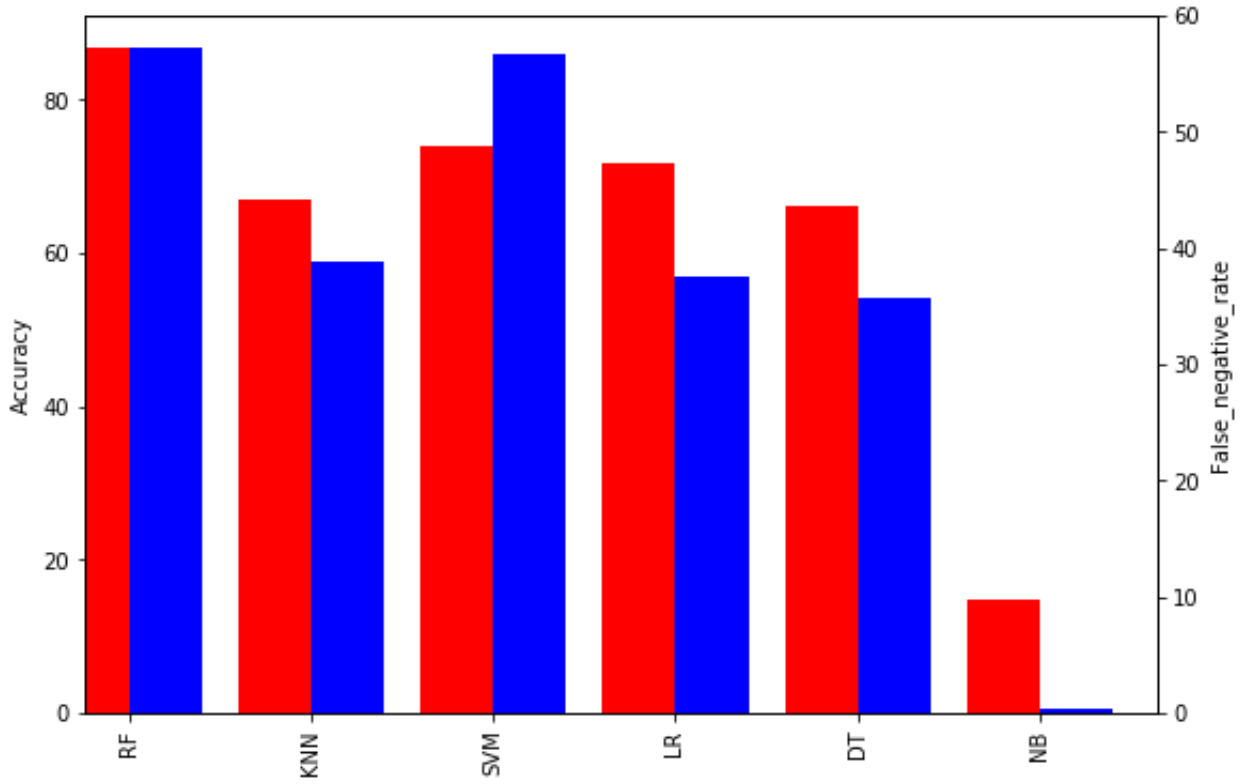



48
46
44
42
40
False_negative_rate
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Number_of_points

3 Chapter 3
Conclusion

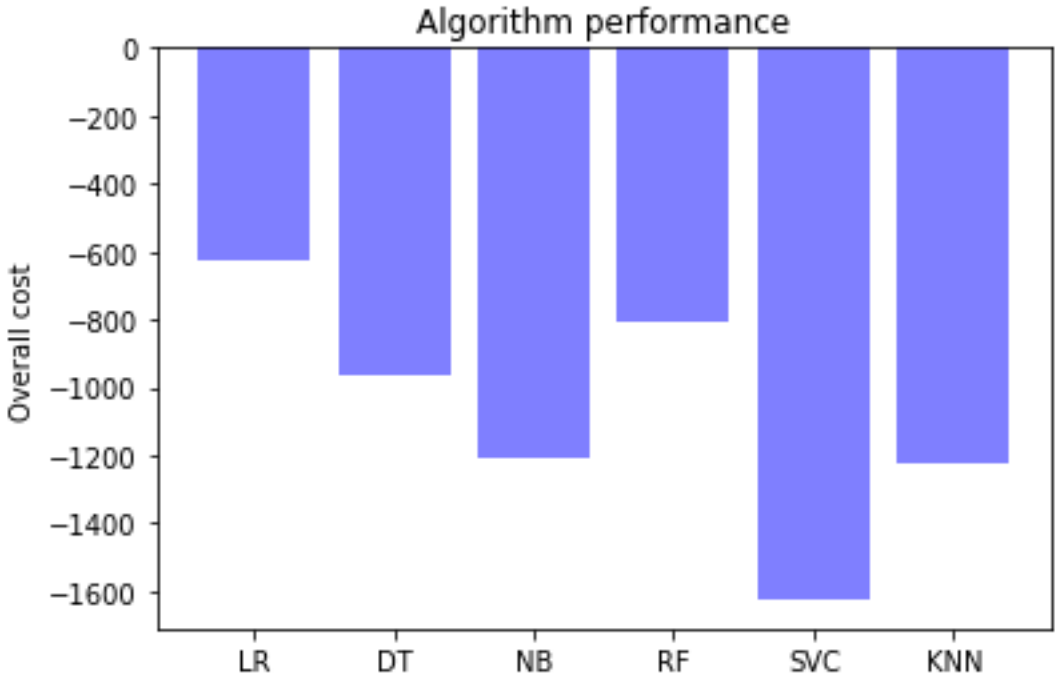
3.1 Error Metrics

As discussed before false negative rate along with accuracy has been chosen as error metrics. But there is also confusion as there are different model giving us different accuracies and false negative rate. The situation can be best described from the following figure



3.2 Cost Optimization and model selection

As error metrics failed to present a clear picture on which model need to be chosen, it is decided to do a cost optimization. In case of business scenario, we assume if a customer is going to churn out, company should reach out to the customer to retain them. So let's assume reaching out to the customer cost 2$ and retaining them gains 10$ for each of them, so in case of each true positive company is gaining 8$ and for each false positive losing 2$. Also in case of false negative company loses 10$ for each customer and for true negative as no reaching out is there so it is 0$. The cost optimization with these assumptions gave us the below presentation from which Logistic regression has been chosen as our final model. In case of R C5 decision tree model gave us the best performance. The representation can be found in appendix



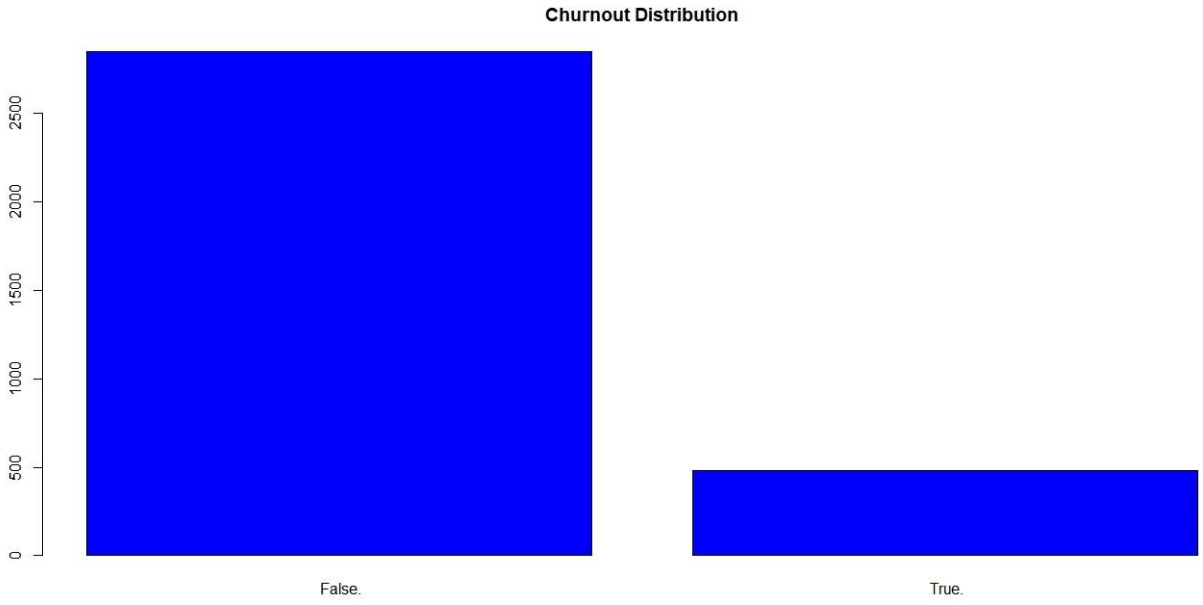
3.3 Model Deployment

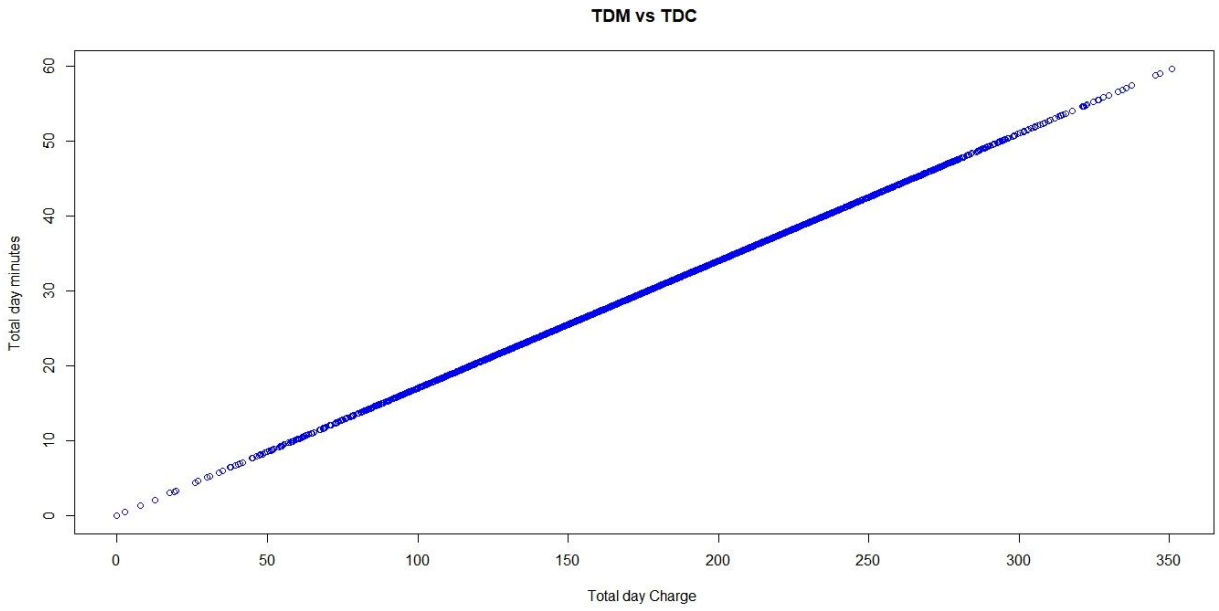
The model is such that the end user only need to run it from the command prompt. The unnecessary training should not be done over and over again just for the prediction purpose. The churnout_main.py file takes care of this. The final model is saved into disk and loaded at the time of prediction. The "Submission_data.csv" file used for this purpose and "Predic_submission.cvs" is our final outcome. In case of R "predict_r_submission.csv" is the final outcome. In case of churnout_predict.R takes care of this particular job.

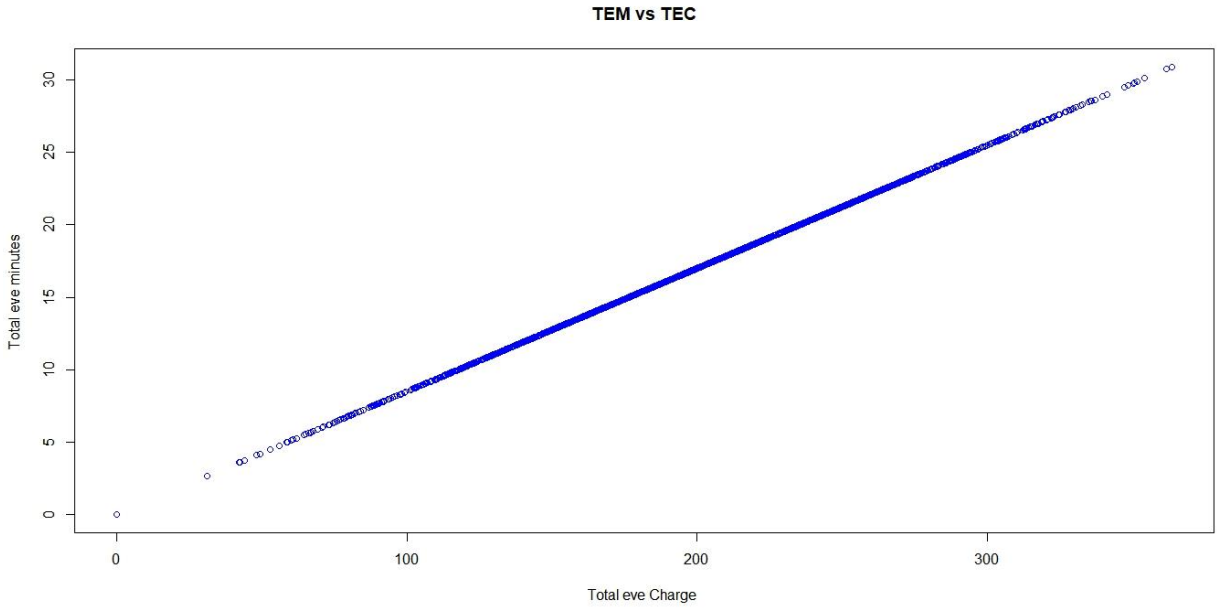
```
(base) C:\Users\Rahul\Desktop\edwisor>Python Churnout_main.py --Data_File C:/Users/Rahul/Desktop/edwisor/Submission_data.csv
Execution started
prediction has been made successfully
```

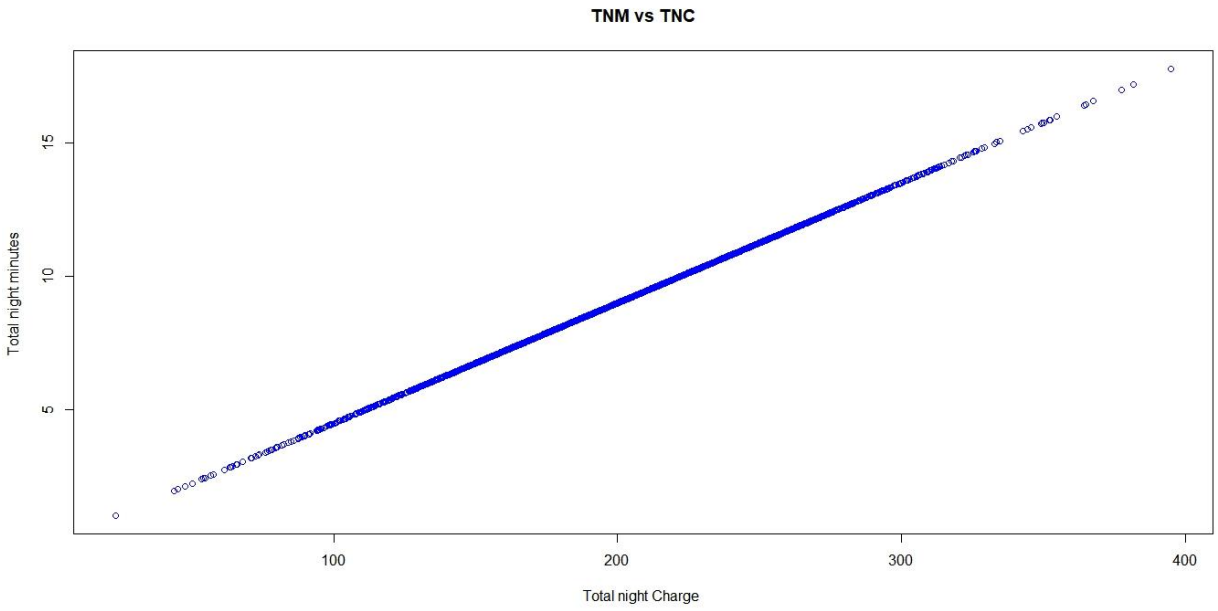
4 Chapter 4
Appendix

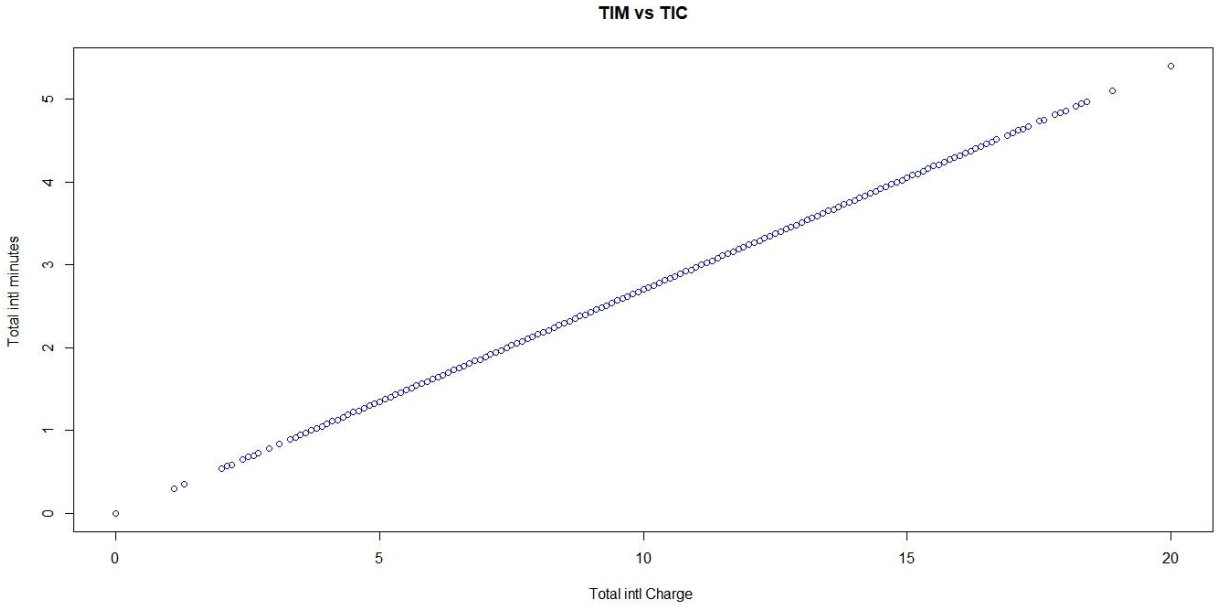
Visualization with R



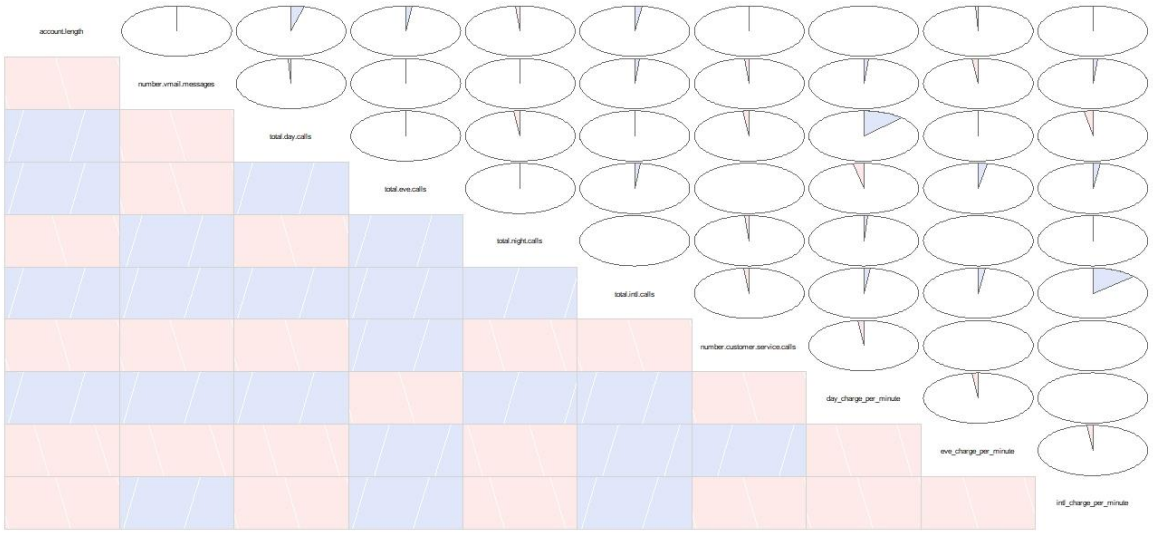

TDM vs TDC
Total day minutes
Total day Charge
0
10
20
30
40
50
60
0
50
100
150
200
250
300
350


TEM vs TEC
Total eve minutes
Total eve Charge
0
5
10
15
20
25
30
0
100
200
300

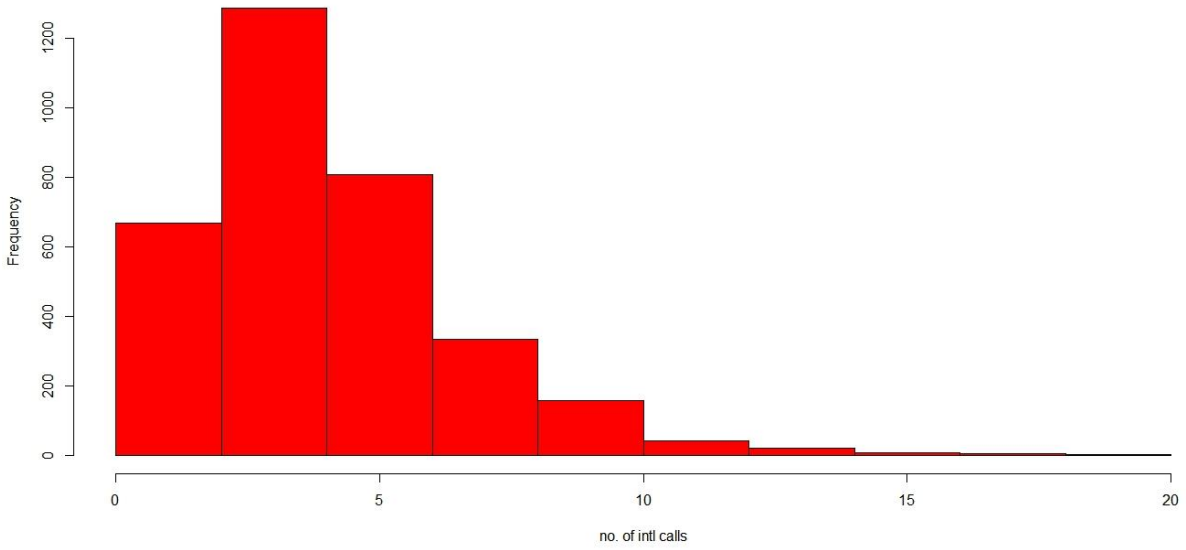

TNM vs TNC
Total night minutes
5
10
15
100
200
300
400
Total night Charge

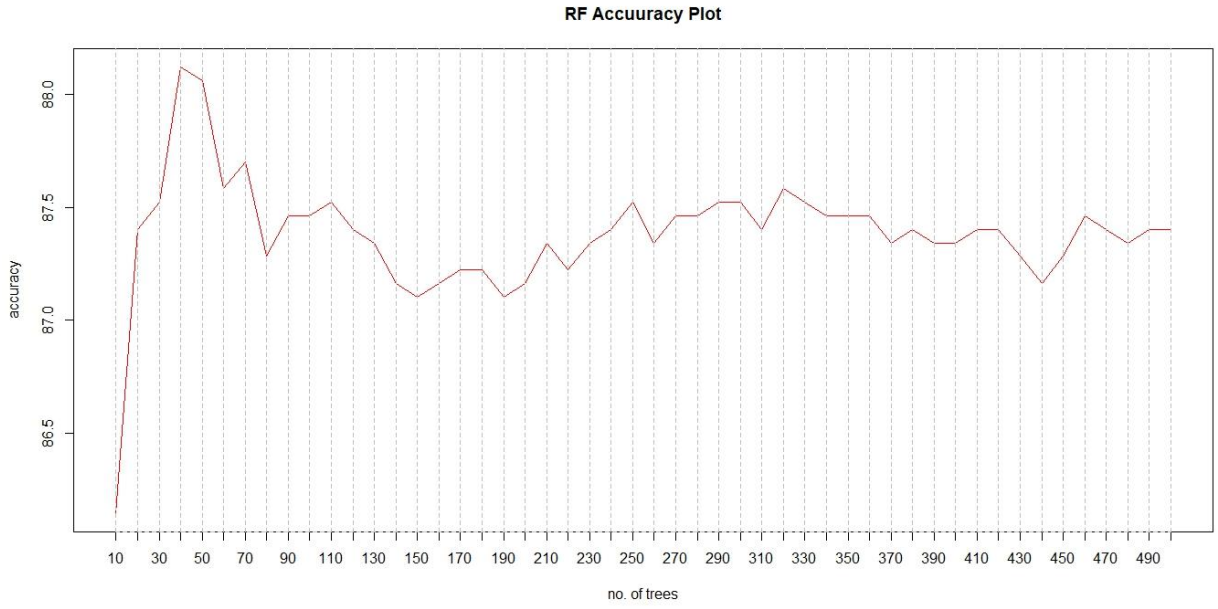

TIM vs TIC
Total intl minutes
0
1
2
3
4
5
0
5
10
15
20
Total intl Charge

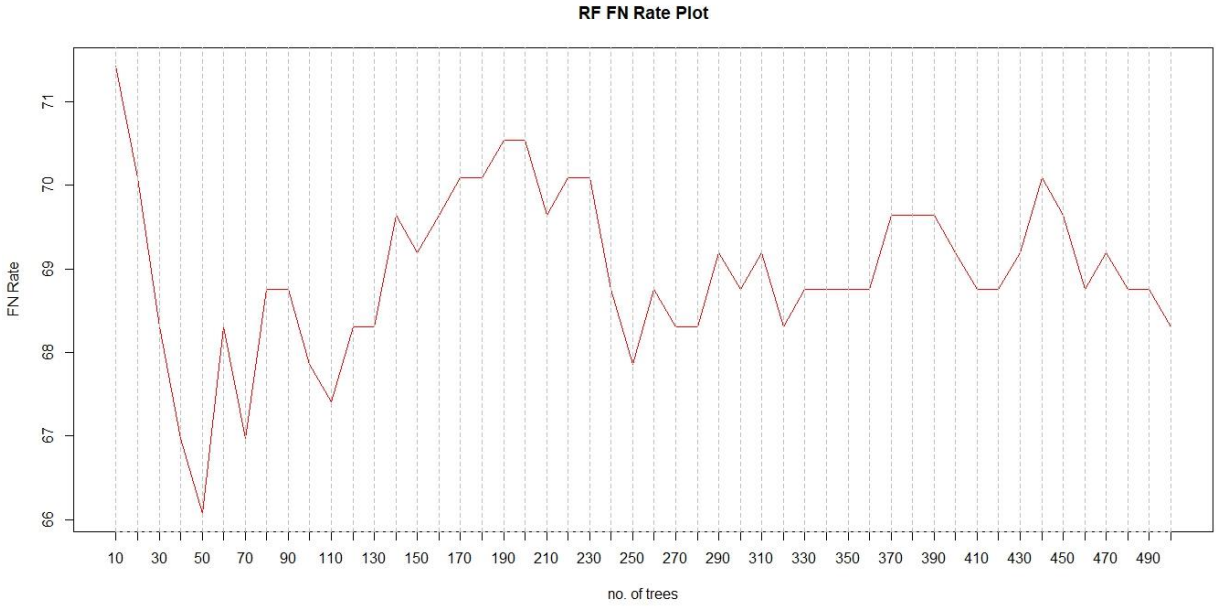
Correlation Plot

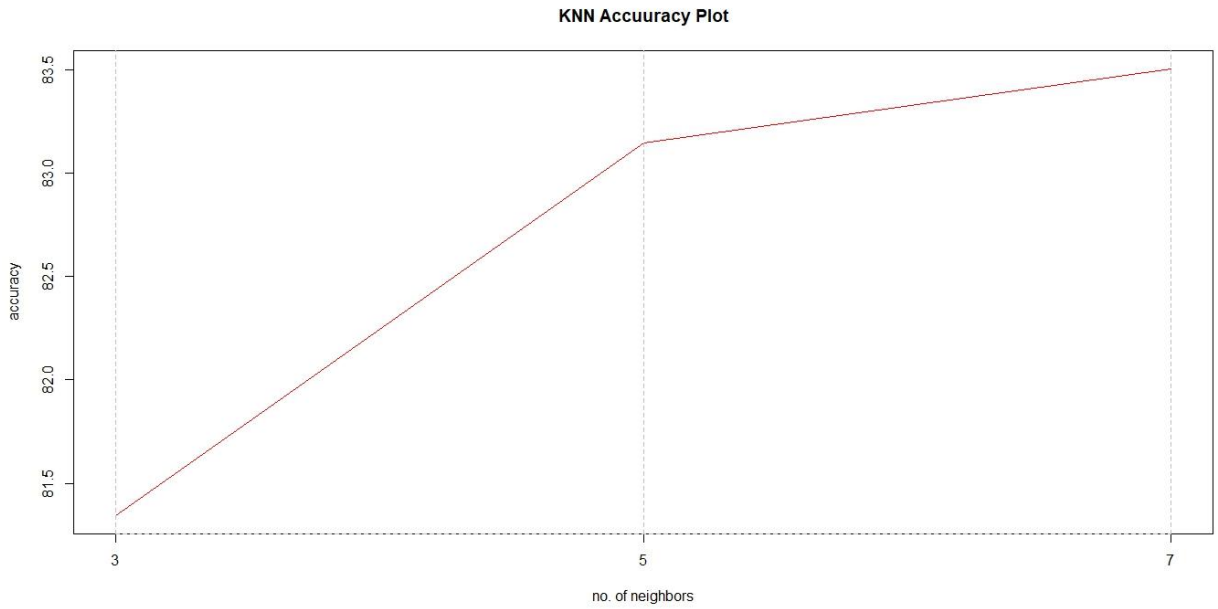


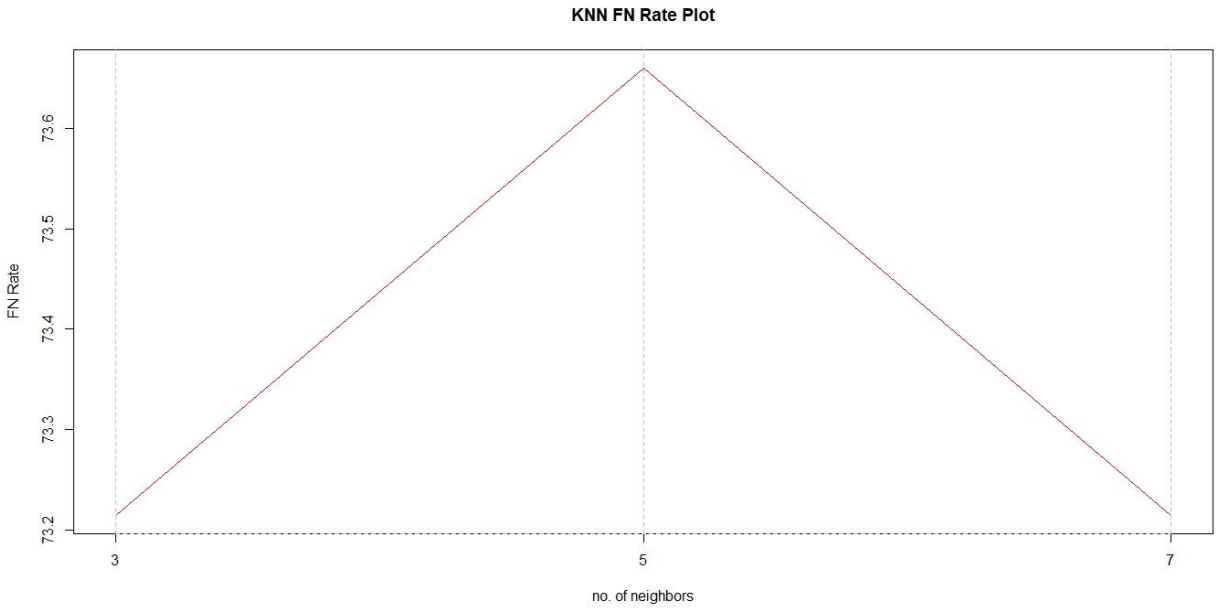
Intl call distribution

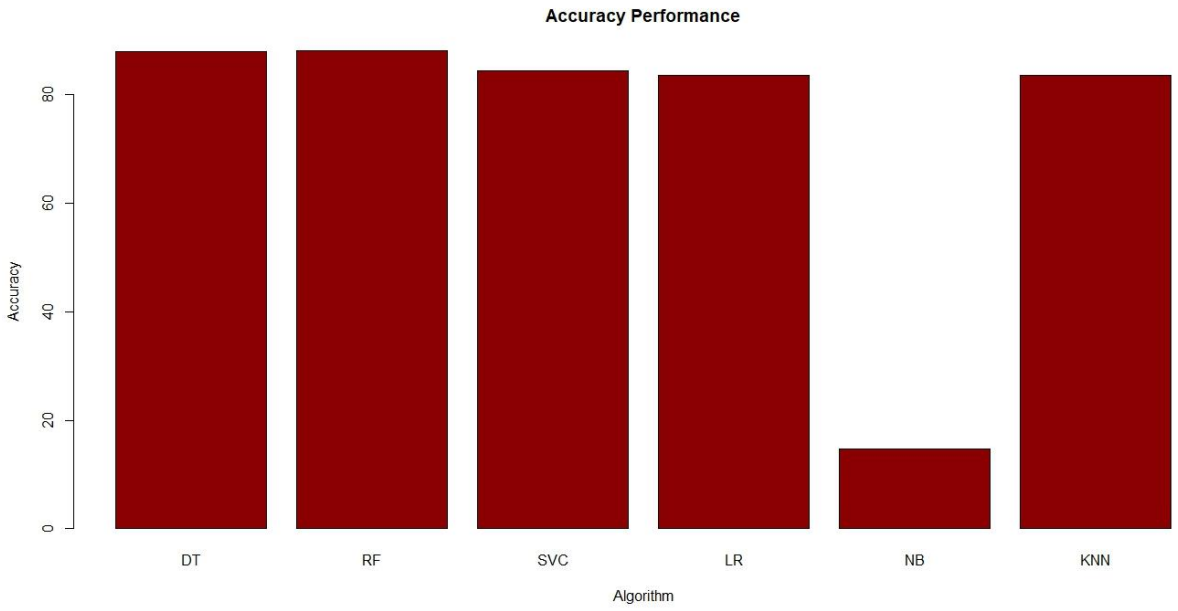


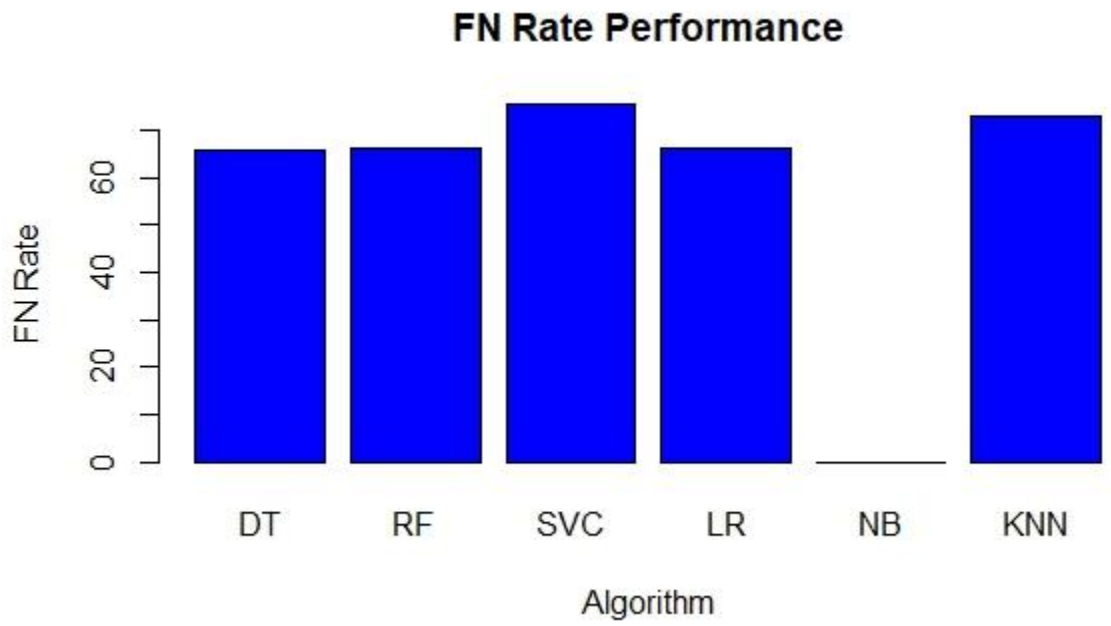

RF Accuuracy Plot
accuracy
88.0
87.5
87.0
86.5
10 30 50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390 410 430 450 470 490
no. of trees

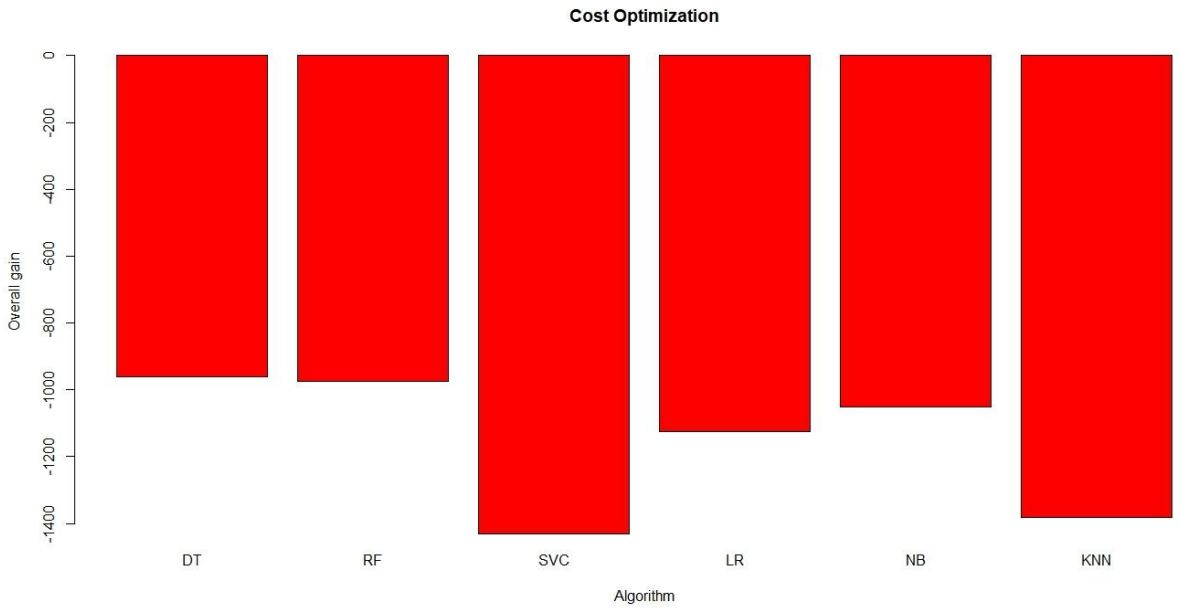

RF FN Rate Plot
FN Rate
71
70
69
68
67
66
10 30 50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390 410 430 450 470 490
no. of trees


KNN Accuuracy Plot
accuracy
83.5
83.0
82.5
82.0
81.5
3
5
7
no. of neighbors


KNN FN Rate Plot
FN Rate
73.6
73.5
73.4
73.3
73.2
3
5
7
no. of neighbors


Accuracy Performance
Accuracy
0
20
40
60
80
DT
RF
SVC
LR
NB
KNN
Algorithm


FN Rate Performance
FN Rate
0
20
40
60
DT
RF
SVC
LR
NB
KNN
Algorithm




```
C:\Users\Rahul>Rscript --vanilla C:/Users/Rahul/Desktop/edwisor/churnout_predict.R C:/Users/Rahul/Desktop/edwisor/Submission_data.csv

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

[1] "Excecution started"
[1] "prediction has been made successfully"
```